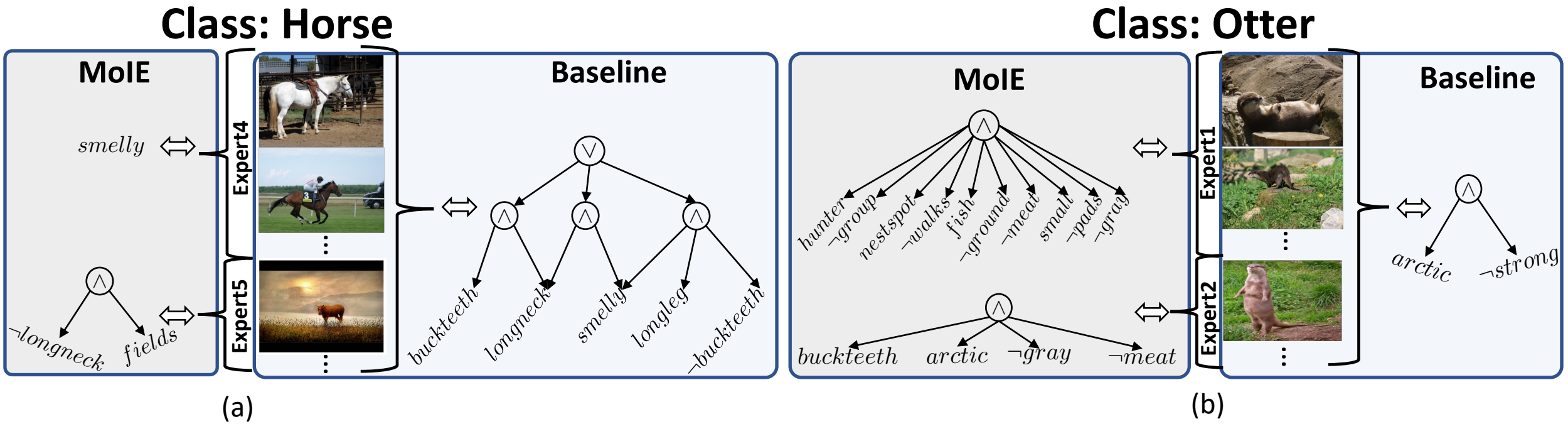
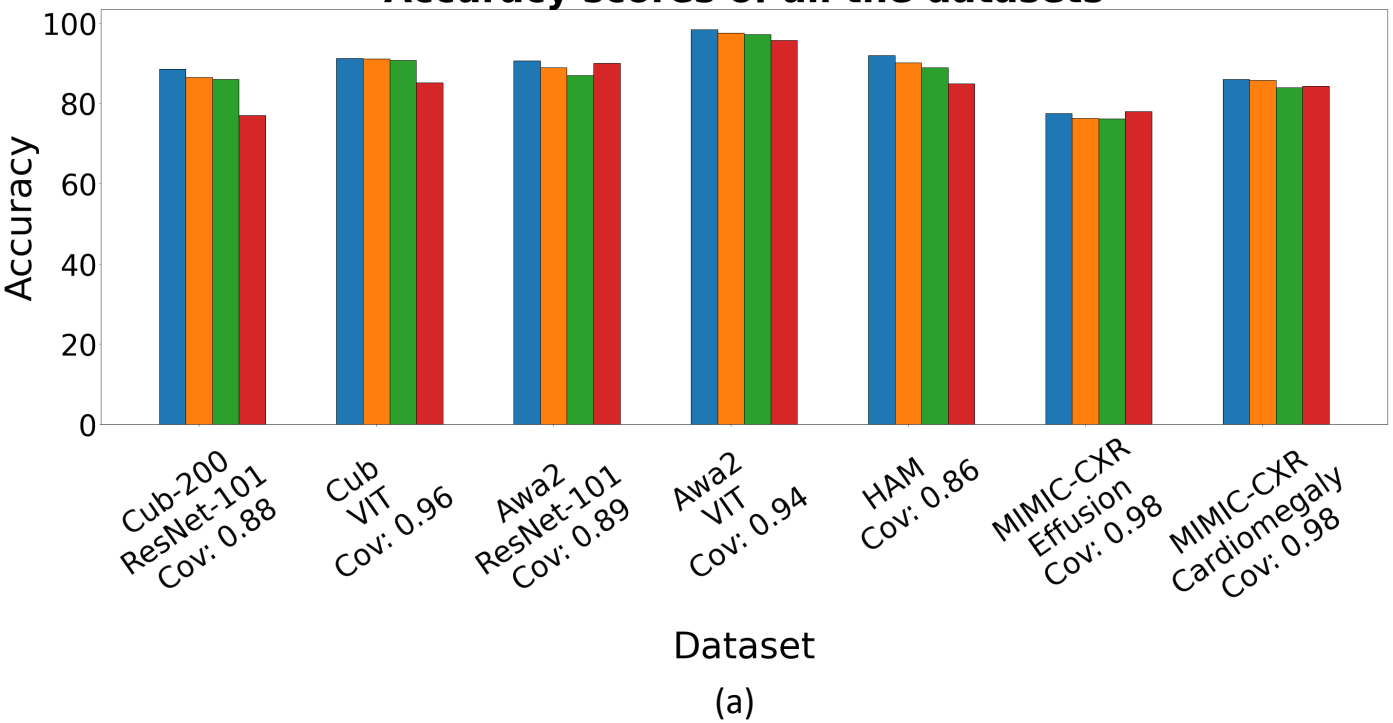
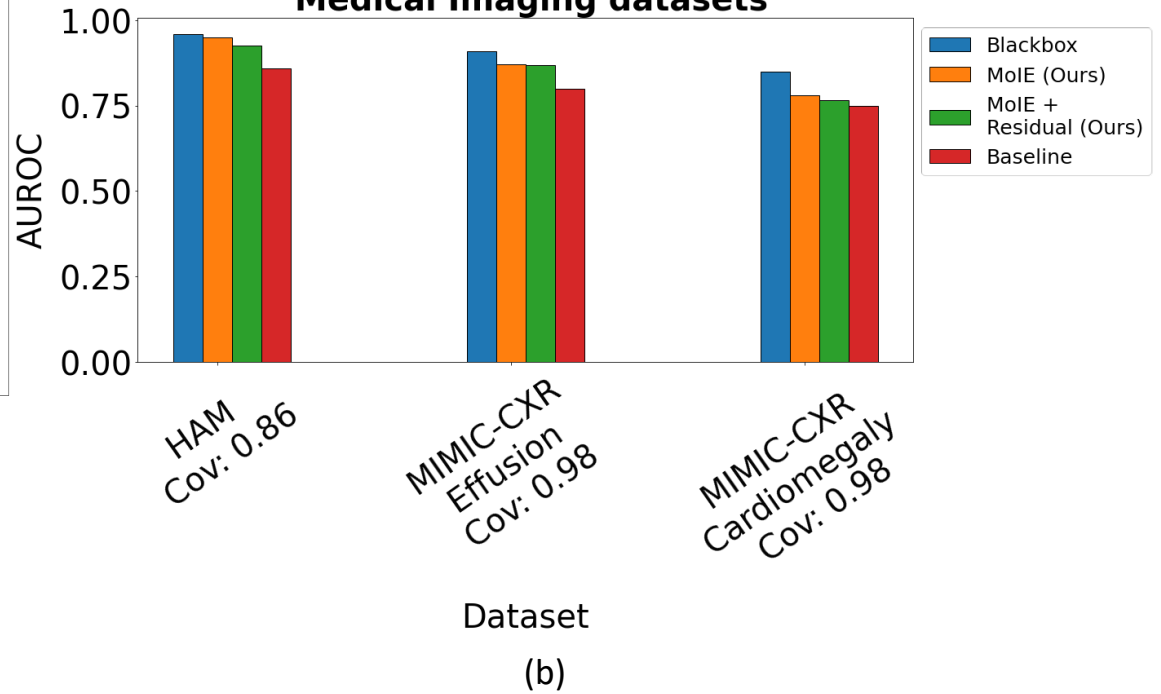
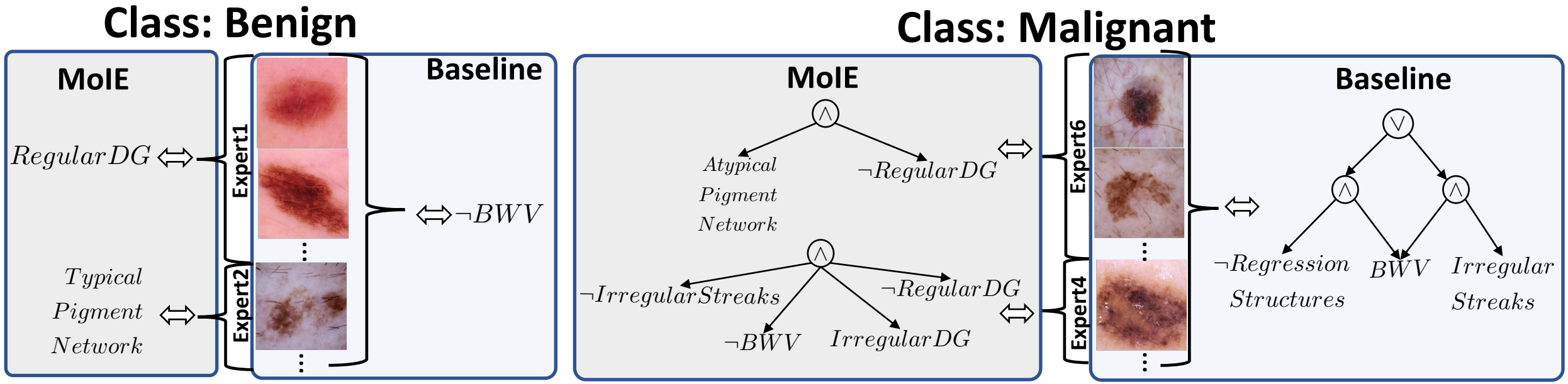

Class: Horse
MoIE
smelly
¬longneck
fields
Expert4
Expert5
Baseline
buckteeth
longneck
smelly
longleg
¬buckteeth
(a)
Class: Otter
MoIE
hunter
¬group
nestspot
¬walks
fish
¬ground
¬meat
small
¬pads
¬gray
buckteeth
arctic
¬gray
¬meat
Expert1
Expert2
Baseline
arctic
¬strong
(b)

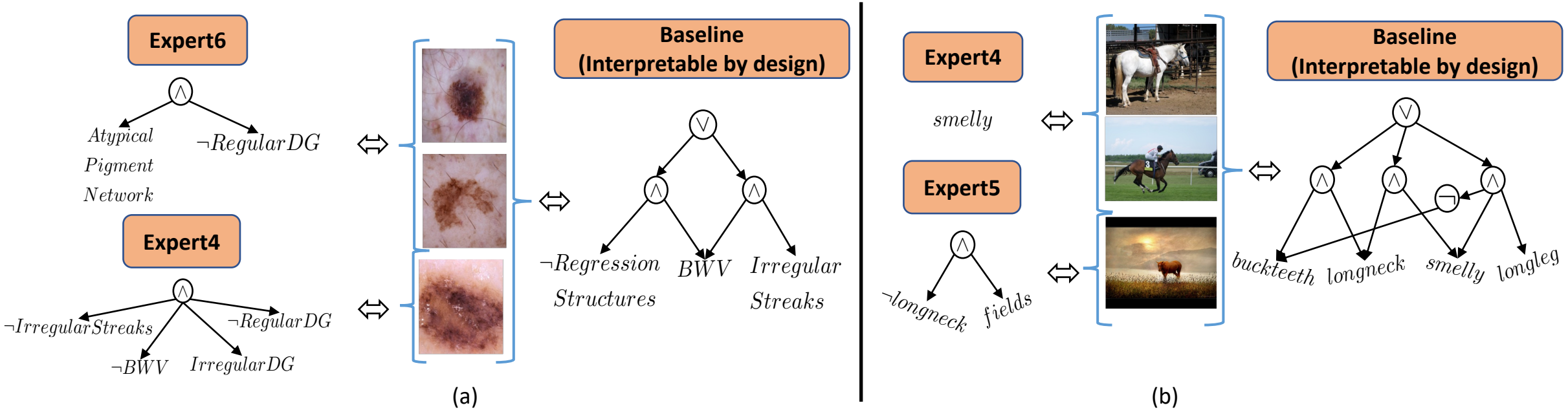

Accuracy scores of all the datasets
Accuracy
100
80
60
40
20
0
Cub-200 ResNet-101 Cov: 0.88
Cub VIT Cov: 0.96
Awa2 ResNet-101 Cov: 0.89
Awa2 VIT Cov: 0.94
HAM Cov: 0.86
MIMIC-CXR Effusion Cov: 0.98
MIMIC-CXR Cardiomegaly Cov: 0.98
Dataset

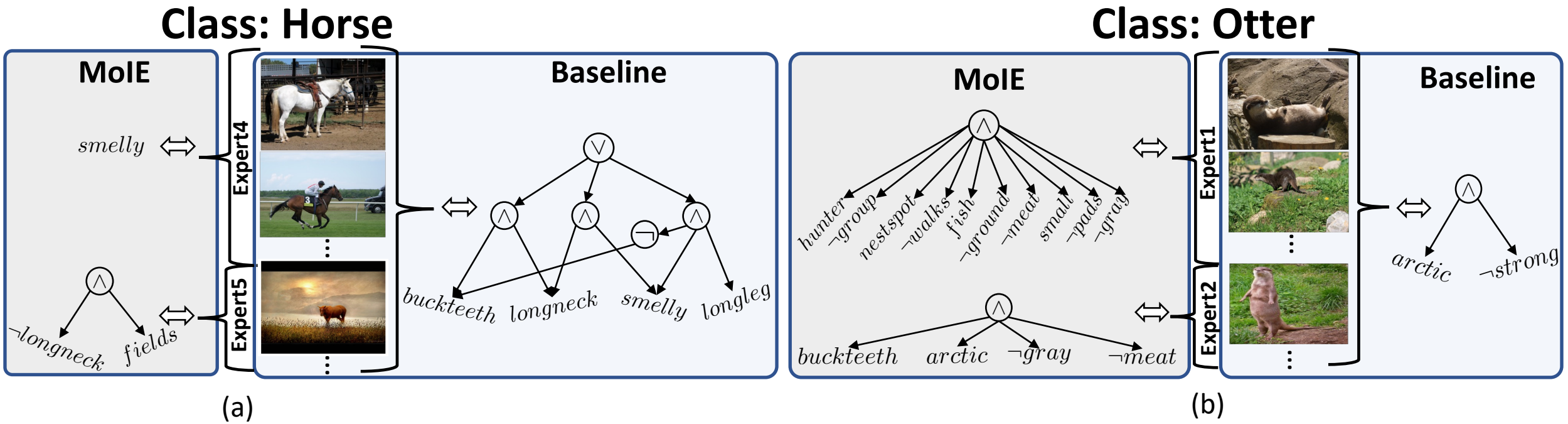
(a)


AUROC scores of the Medical Imaging datasets
AUROC
1.00
0.75
0.50
0.25
0.00
HAM Cov: 0.86
MIMIC-CXR Effusion Cov: 0.98
MIMIC-CXR Cardiomegaly Cov: 0.98
Dataset
Blackbox
MoIE (Ours)
MoIE + Residual (Ours)
Baseline

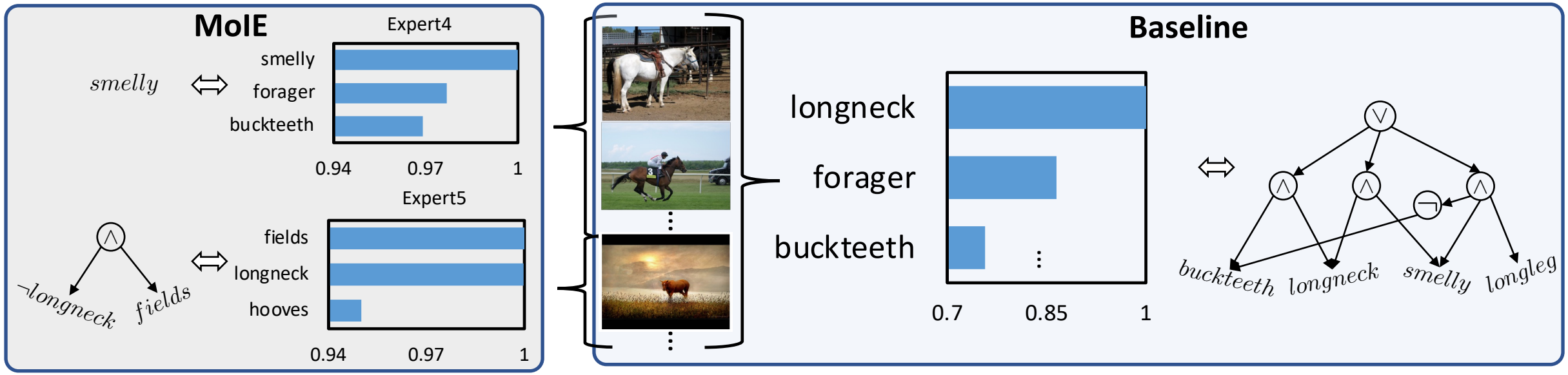
(b)


Class: Benign
MoIE
$RegularDG \Leftrightarrow$
$Typical\ Pigment\ Network \Leftrightarrow$
Expert1
Expert2
Baseline
$\Leftrightarrow \neg BWV$
(c)
Class: Malignant
MoIE
$\wedge$
$Atypical\ Pigment\ Network$
$\neg RegularDG$
$\neg IrregularStreaks$
$\neg RegularDG$
$\neg BWV$
$IrregularDG$
Expert6
Expert4
Baseline
$\vee$
$\wedge$
$\wedge$
$\neg Regression\ Structures$
BWV
$Irregular\ Streaks$
(d)

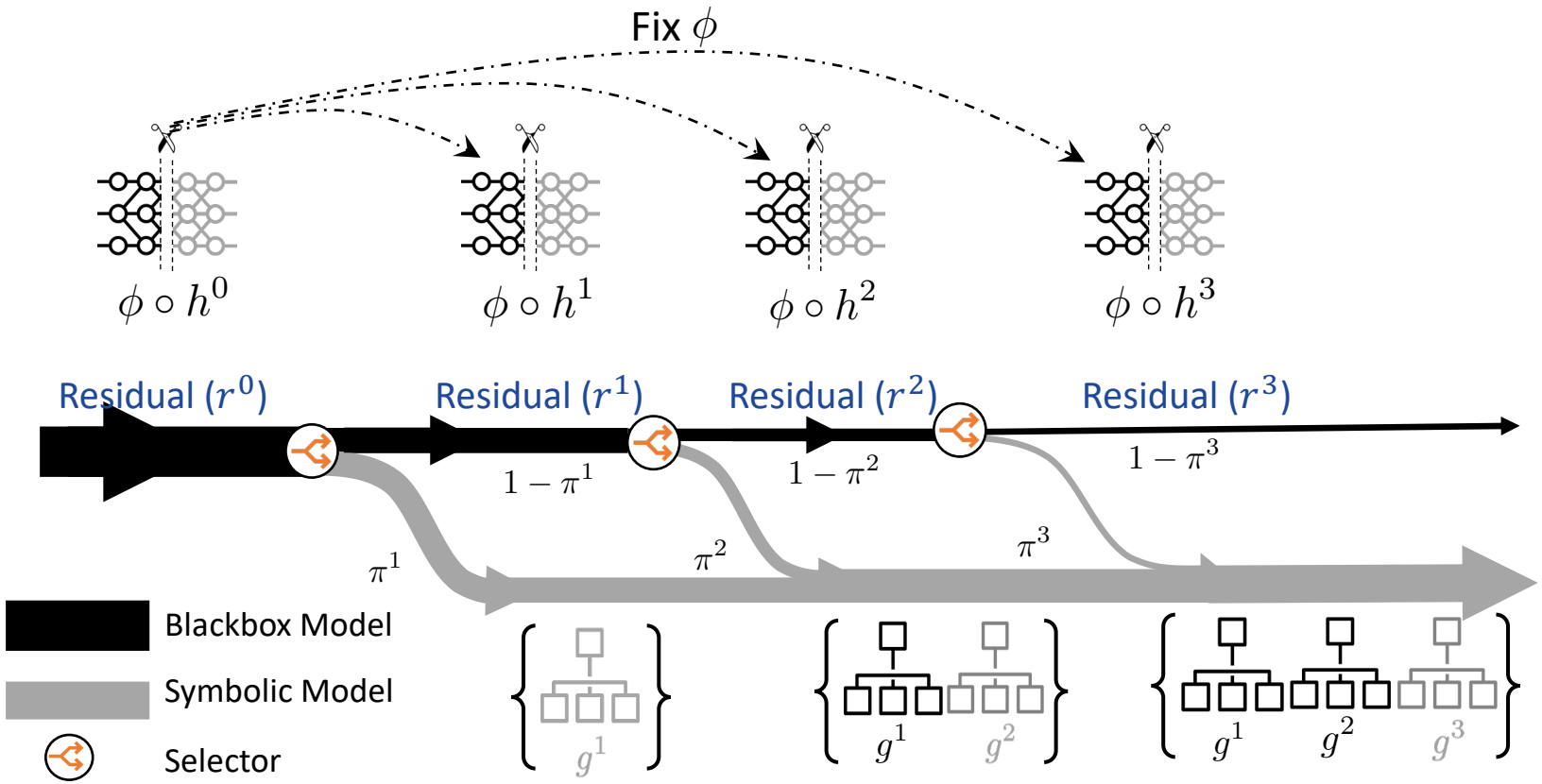

Expert6
Atypical Pigment Network
¬RegularDG
Expert4
¬IrregularStreaks
¬RegularDG
¬BWV
IrregularDG
Baseline (Interpretable by design)
¬Regression Structures
BWV
Irregular Streaks
(a)
Expert4
smelly
Expert5
¬longneck
fields
Baseline (Interpretable by design)
buckteeth
longneck
smelly
longleg
(b)


Class: Horse
MoIE
smelly
¬longneck
fields
Expert4
Expert5
Baseline
buckteeth
longneck
smelly
longleg
(a)
Class: Otter
MoIE
hunter
¬group
nestspot
¬walks
fish
¬ground
¬meat
small
¬pads
¬gray
buckteeth
arctic
¬gray
¬meat
Expert1
Expert2
Baseline
arctic
¬strong
(b)

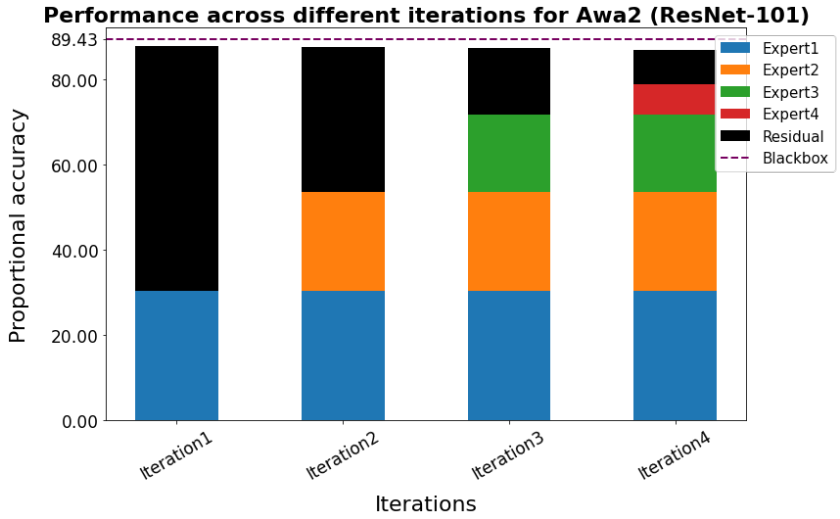
Class: Horse



(a)

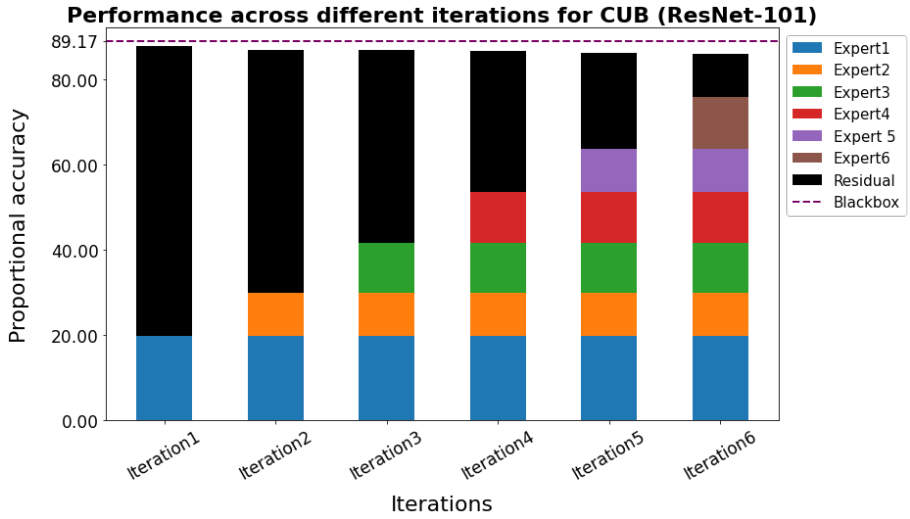

Fix ϕ
$\phi \circ h^0$
$\phi \circ h^1$
$\phi \circ h^2$
$\phi \circ h^3$
Residual (r^0)
Residual (r^1)
Residual (r^2)
Residual (r^3)
$1-\pi^1$
$1-\pi^2$
$1-\pi^3$
π^1
π^2
π^3
Blackbox Model
Symbolic Model
Selector
g^1
g^1
g^2
g^1
g^2
g^3


Coverages of various experts for Awa2 (ResNet-101)
0.0
0.2
0.4
0.6
0.8
1.0

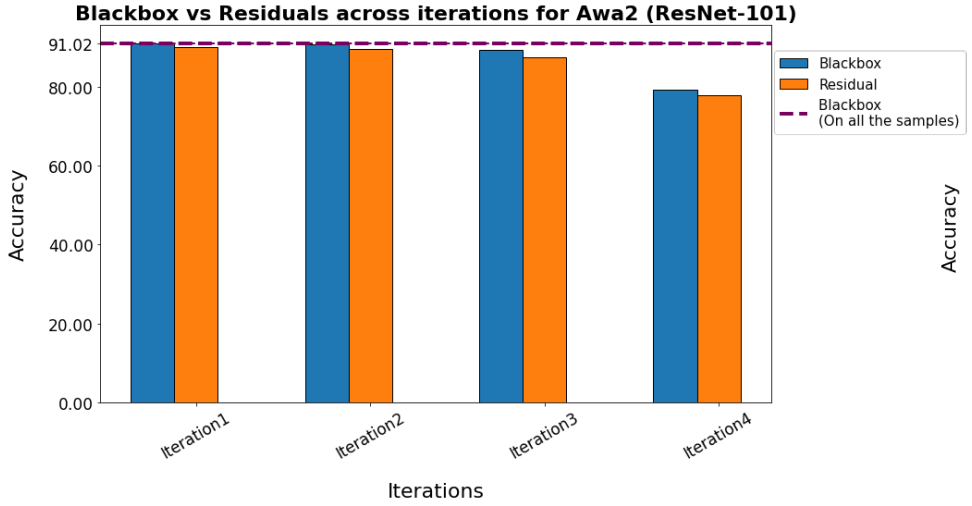

Performance across different iterations for Awa2 (ResNet-101)
89.43
80.00
60.00
40.00
20.00
0.00
Proportional accuracy
Iteration1
Iteration2
Iteration3
Iteration4
Iterations
Expert1
Expert2
Expert3
Expert4
Residual
Blackbox

(a)

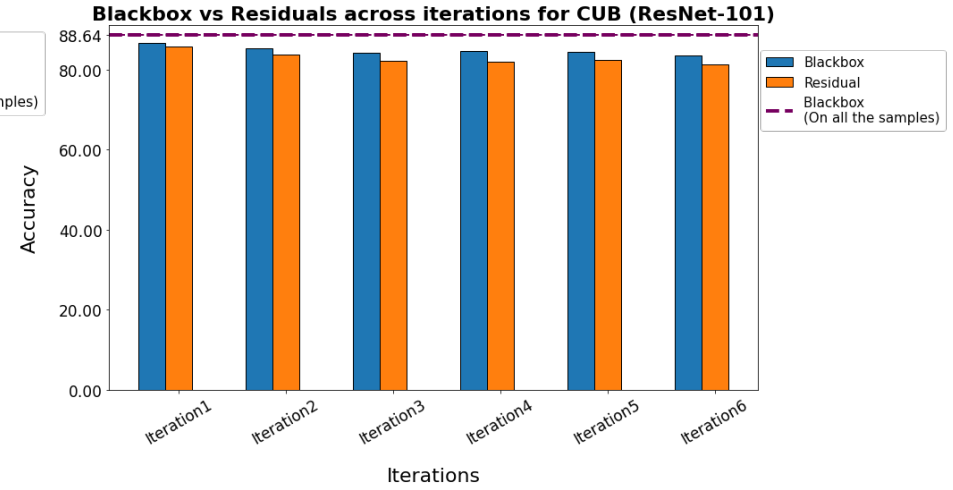

Coverages of various experts for CUB (ResNet-101)
0.0
0.2
0.4
0.6
0.8
1.0


Performance across different iterations for CUB (ResNet-101)
89.17
80.00
60.00
40.00
20.00
0.00
Proportional accuracy
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations
Expert1
Expert2
Expert3
Expert4
Expert 5
Expert6
Residual
Blackbox

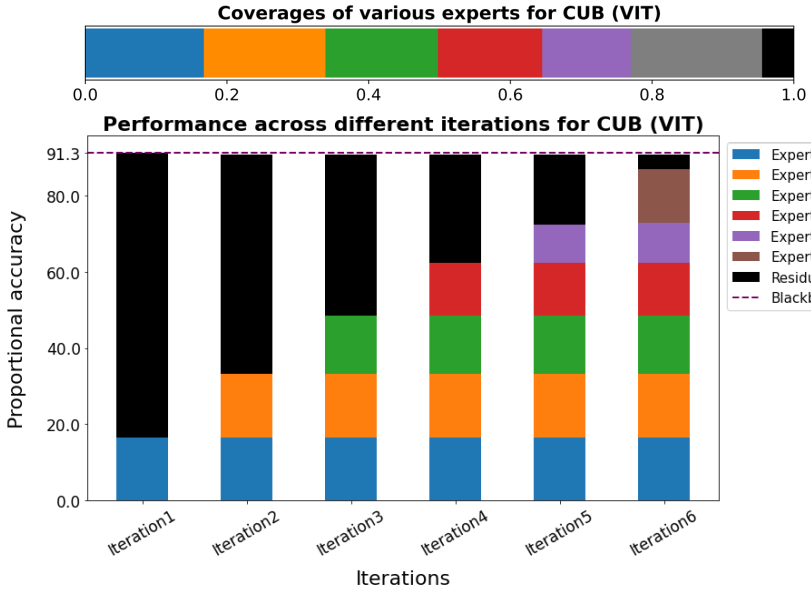
(b)


Blackbox vs Residuals across iterations for Awa2 (ResNet-101)
91.02
80.00
60.00
40.00
20.00
0.00
Accuracy
Iteration1
Iteration2
Iteration3
Iteration4
Iterations
Blackbox
Residual
Blackbox
(On all the samples)

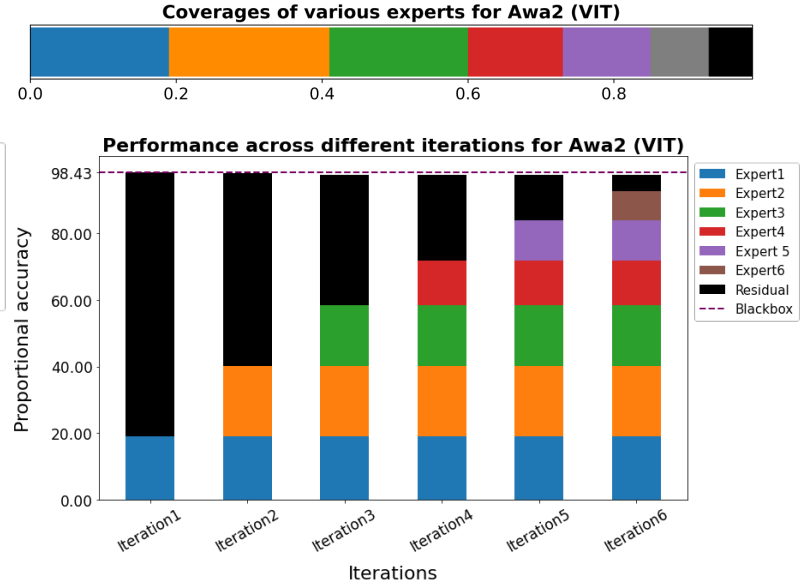
(c)


Blackbox vs Residuals across iterations for CUB (ResNet-101)
88.64
80.00
60.00
40.00
20.00
0.00
Accuracy
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations
Blackbox
Residual
Blackbox
(On all the samples)

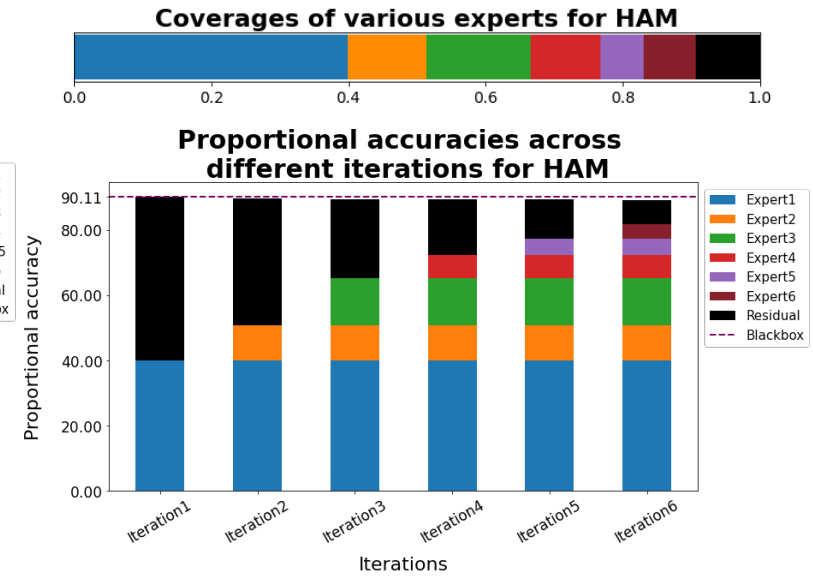
(d)


Coverages of various experts for CUB (VIT)
0.0
0.2
0.4
0.6
0.8
1.0
Performance across different iterations for CUB (VIT)
Proportional accuracy
91.3
80.0
60.0
40.0
20.0
0.0
Expert1
Expert2
Expert3
Expert4
Expert 5
Expert6
Residual
Blackbox
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations

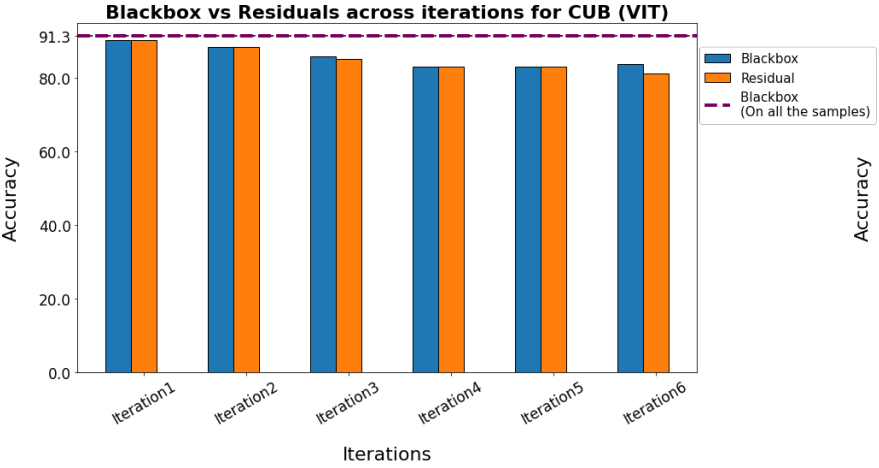
(a)


Coverages of various experts for Awa2 (VIT)
0.0
0.2
0.4
0.6
0.8
Performance across different iterations for Awa2 (VIT)
Proportional accuracy
98.43
80.00
60.00
40.00
20.00
0.00
Expert1
Expert2
Expert3
Expert4
Expert 5
Expert6
Residual
Blackbox
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations

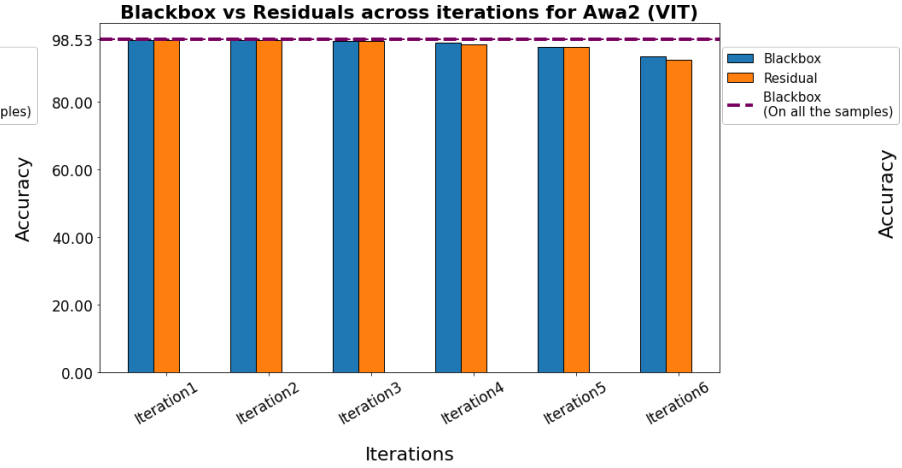
(b)


Coverages of various experts for HAM
0.0
0.2
0.4
0.6
0.8
1.0
Proportional accuracies across different iterations for HAM
Proportional accuracy
90.11
80.00
60.00
40.00
20.00
0.00
Expert1
Expert2
Expert3
Expert4
Expert5
Expert6
Residual
Blackbox
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations

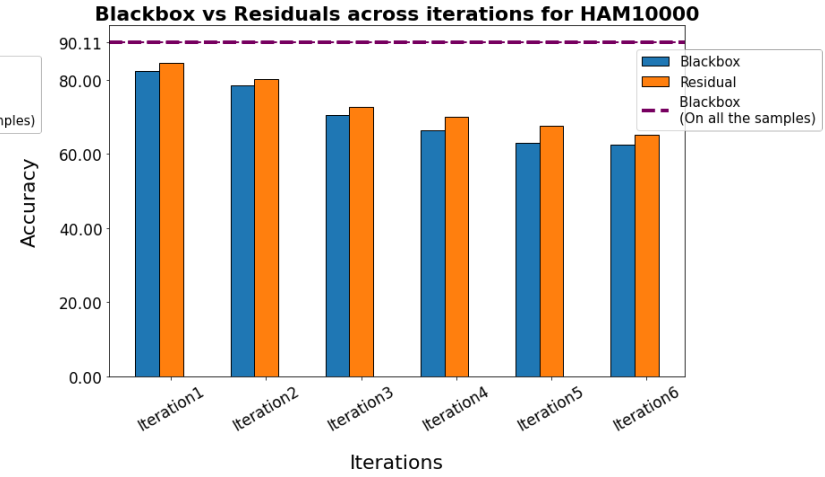
(c)


Blackbox vs Residuals across iterations for CUB (VIT)
Accuracy
91.3
80.0
60.0
40.0
20.0
0.0
Blackbox
Residual
Blackbox
(On all the samples)
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations

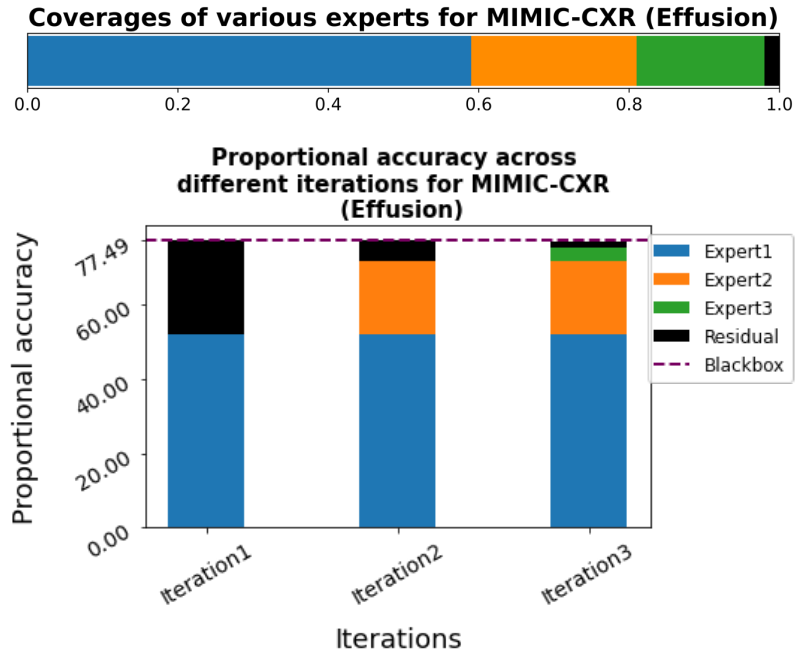
(d)


Blackbox vs Residuals across iterations for Awa2 (VIT)
Accuracy
98.53
80.00
60.00
40.00
20.00
0.00
Blackbox
Residual
Blackbox
(On all the samples)
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations

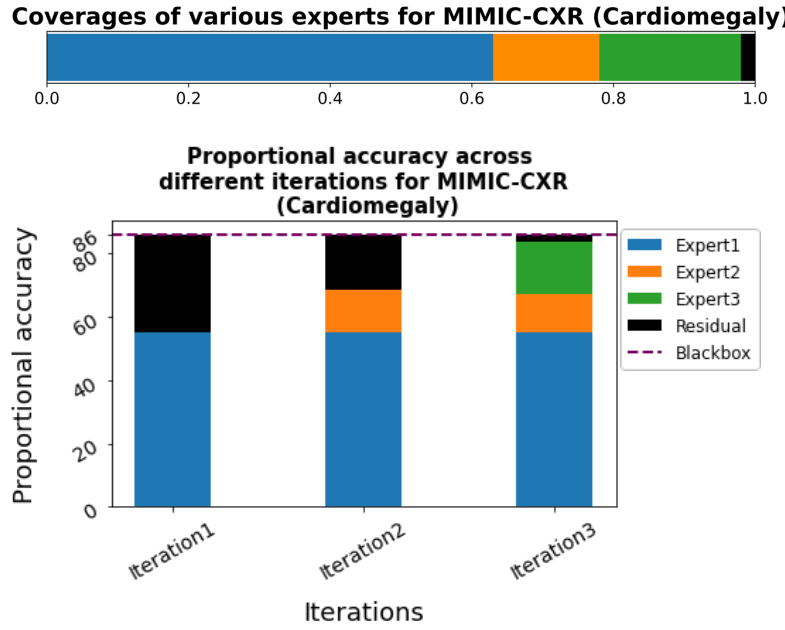
(e)


Blackbox vs Residuals across iterations for HAM10000
Accuracy
90.11
80.00
60.00
40.00
20.00
0.00
Blackbox
Residual
Blackbox
(On all the samples)
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations

(f)


Coverages of various experts for MIMIC-CXR (Effusion)
0.0
0.2
0.4
0.6
0.8
1.0
Proportional accuracy across different iterations for MIMIC-CXR (Effusion)
Proportional accuracy
77.49
60.00
40.00
20.00
0.00
Iteration1
Iteration2
Iteration3
Iterations
Expert1
Expert2
Expert3
Residual
Blackbox

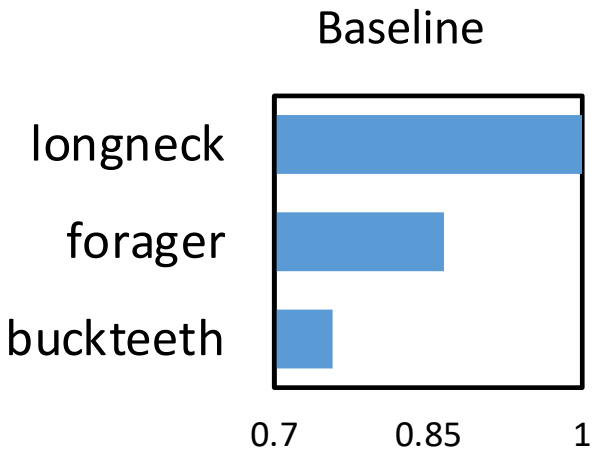
(a)

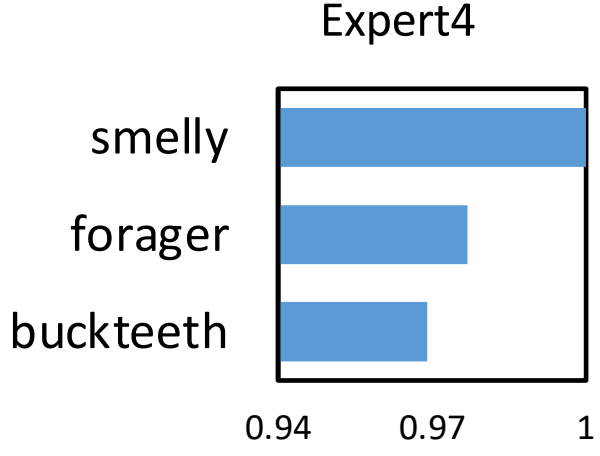

Coverages of various experts for MIMIC-CXR (Cardiomegaly)
0.0
0.2
0.4
0.6
0.8
1.0
Proportional accuracy across different iterations for MIMIC-CXR (Cardiomegaly)
Proportional accuracy
86
80
60
40
20
0
Iteration1
Iteration2
Iteration3
Iterations
Expert1
Expert2
Expert3
Residual
Blackbox

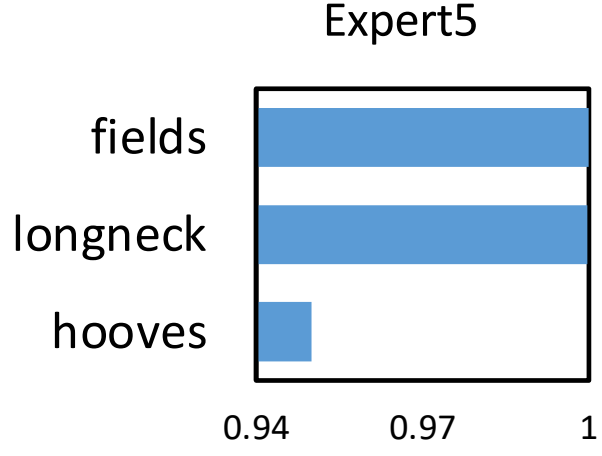
(b)

(a)

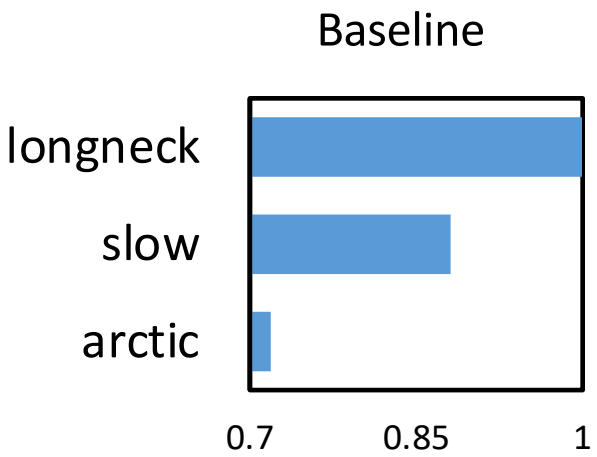
Horse

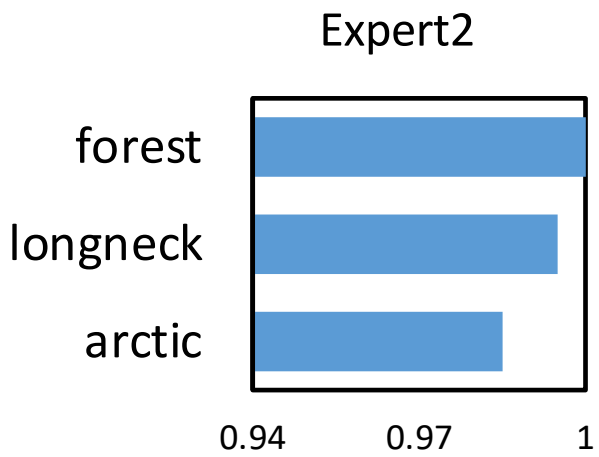

Baseline
longneck
forager
buckteeth
0.7
0.85
1

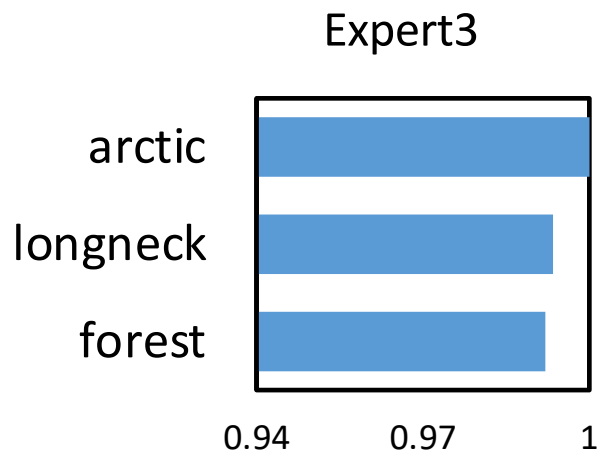

Expert4
smelly
forager
buckteeth
0.94
0.97
1


Expert5
fields
longneck
hooves
0.94
0.97
1

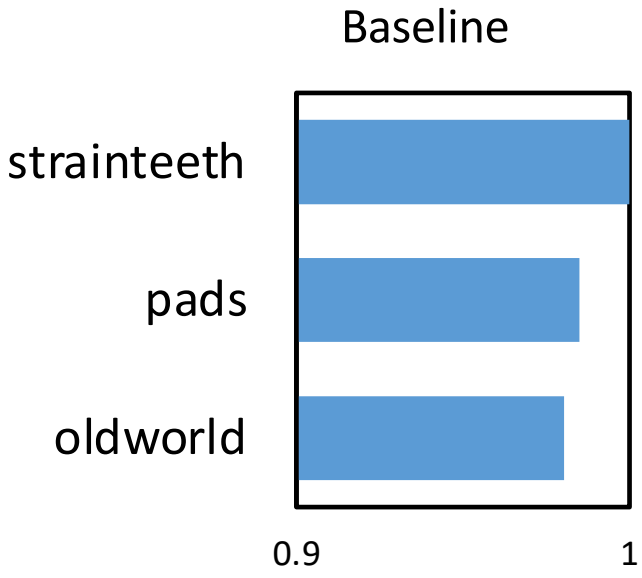
Moose

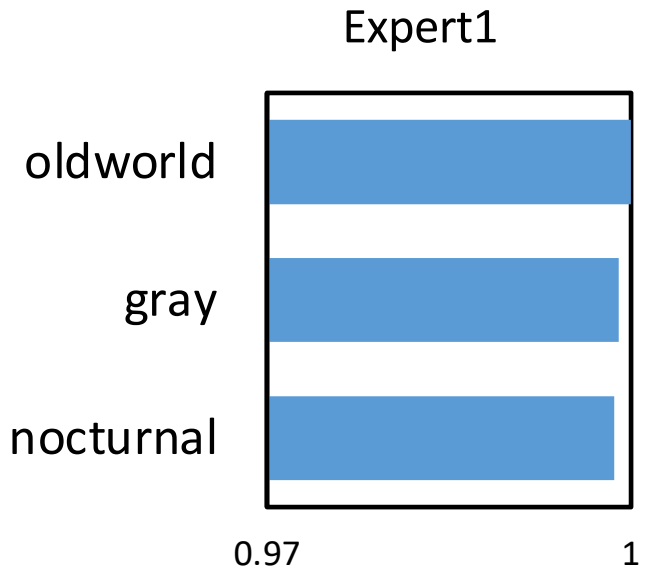

Baseline
longneck
slow
arctic
0.7
0.85
1

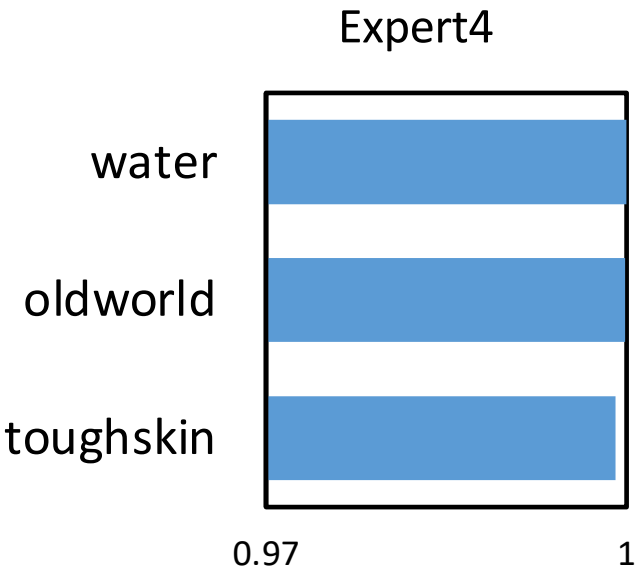

Expert2
forest
longneck
arctic
0.94
0.97
1


Expert3
arctic
longneck
forest
0.94
0.97
1

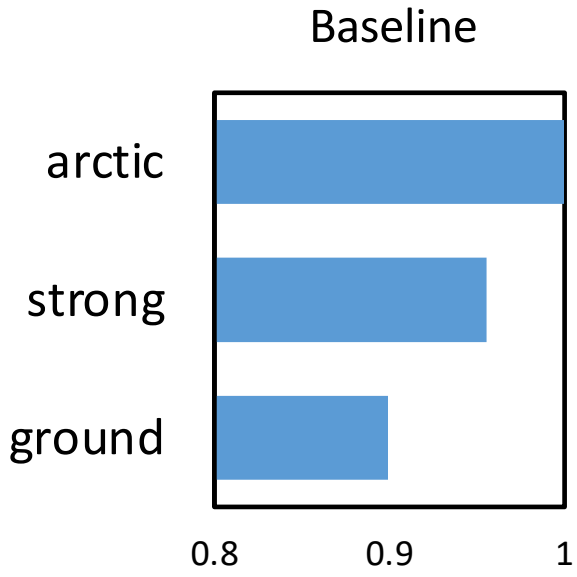
Beaver

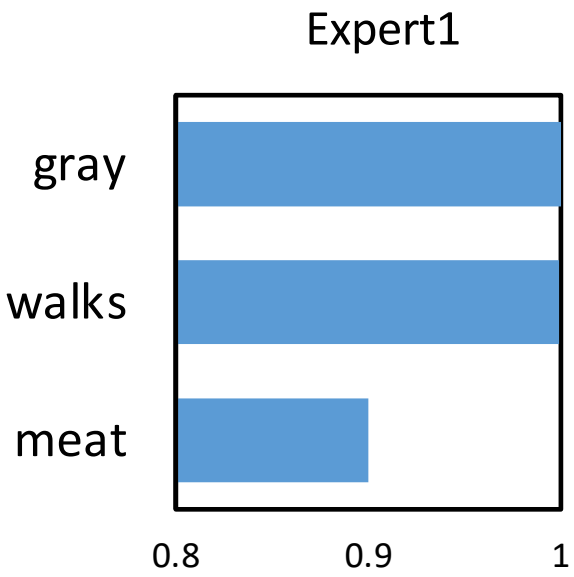




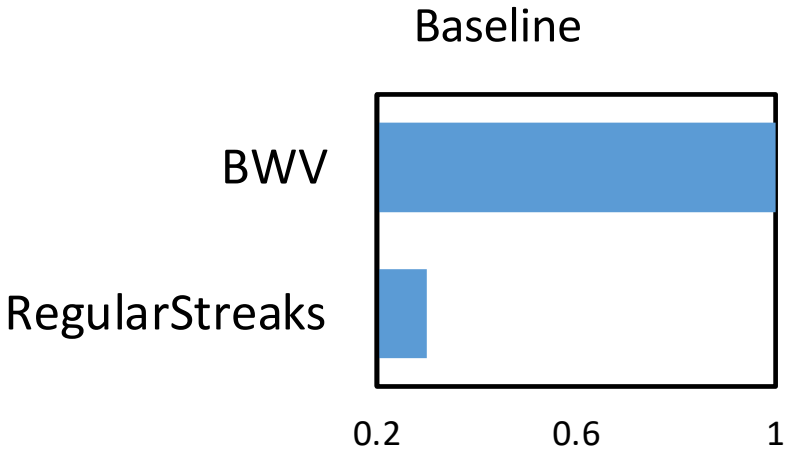


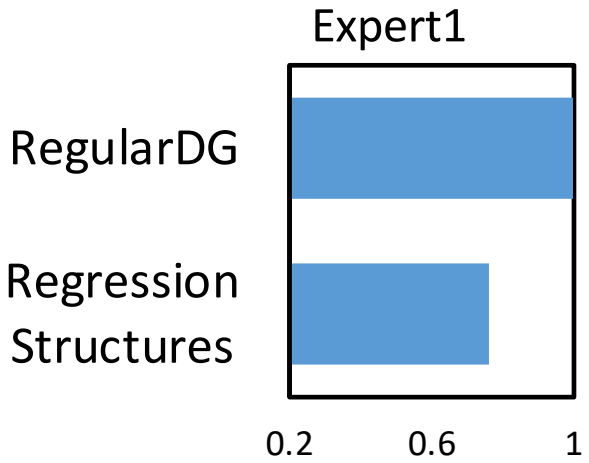
Otter

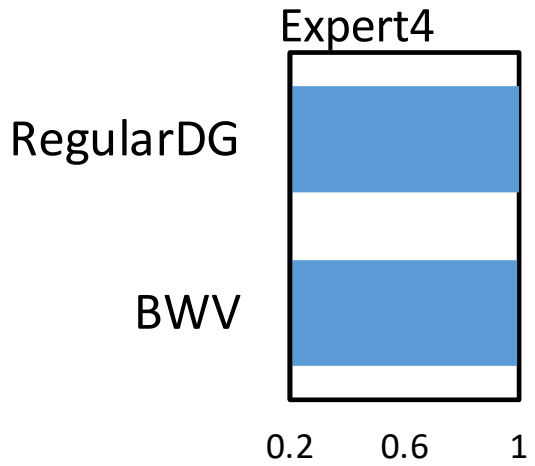




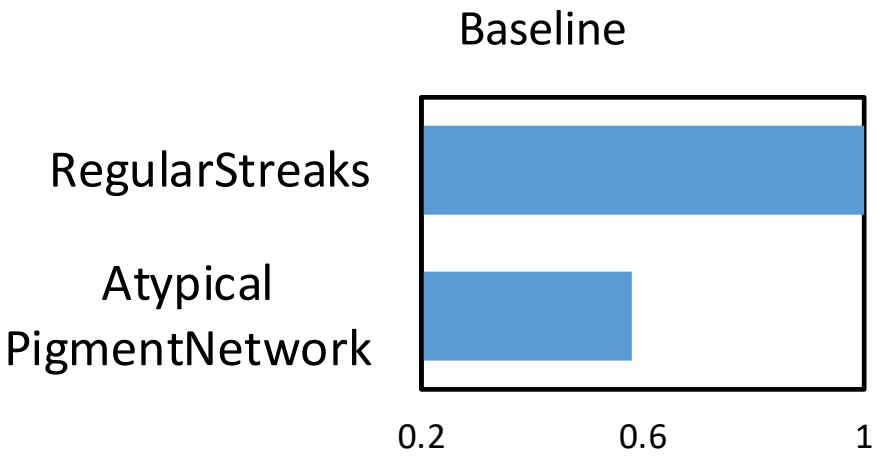
Benign

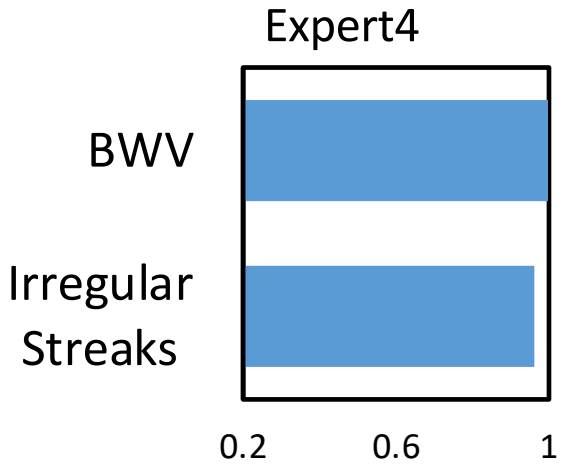


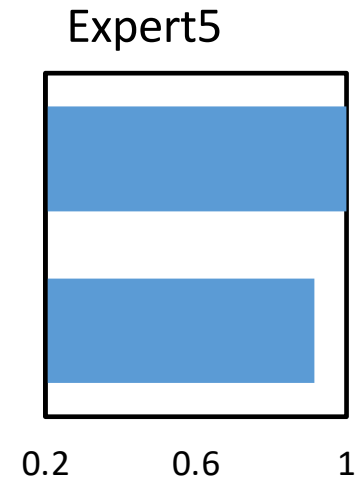


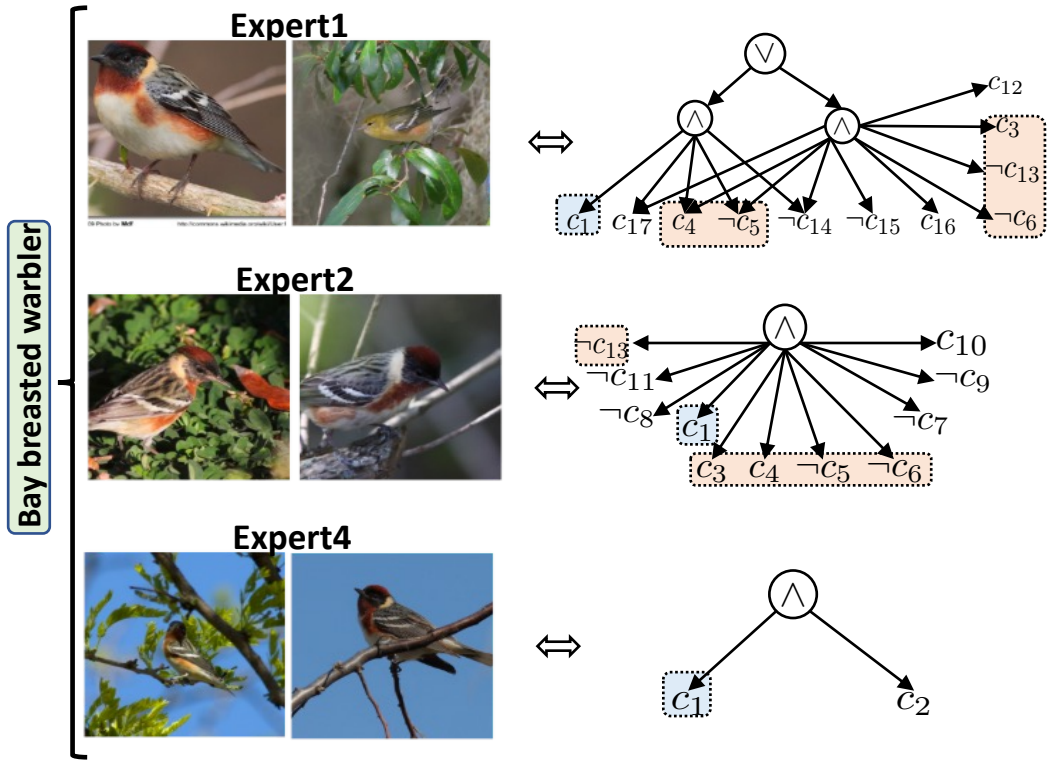


Malignant





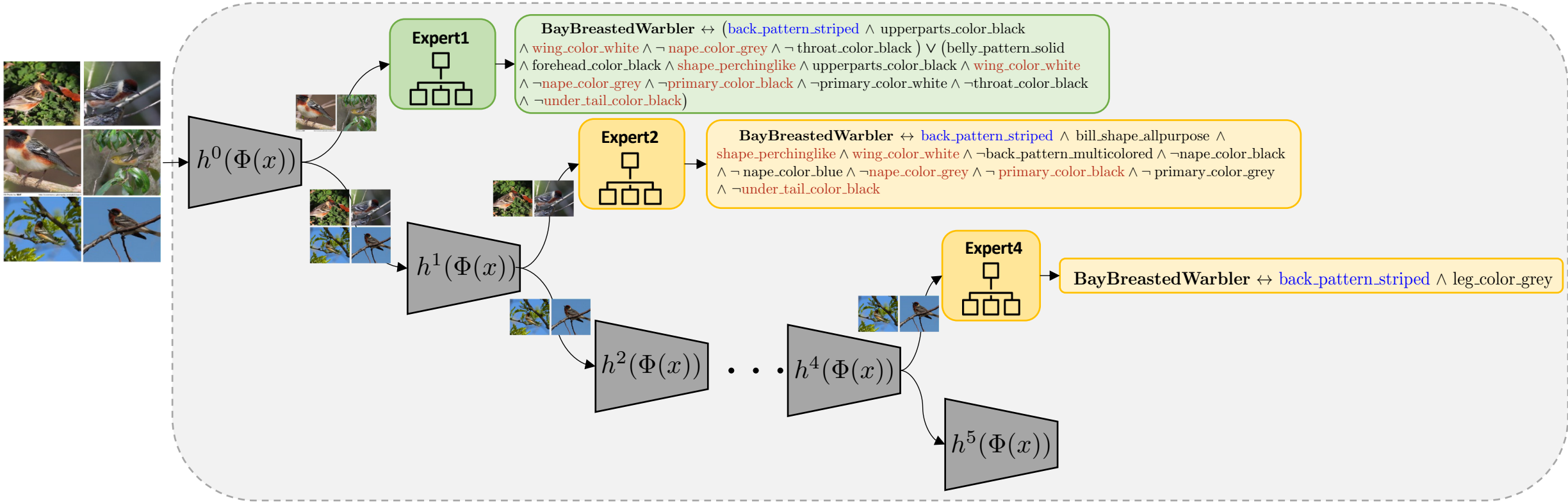


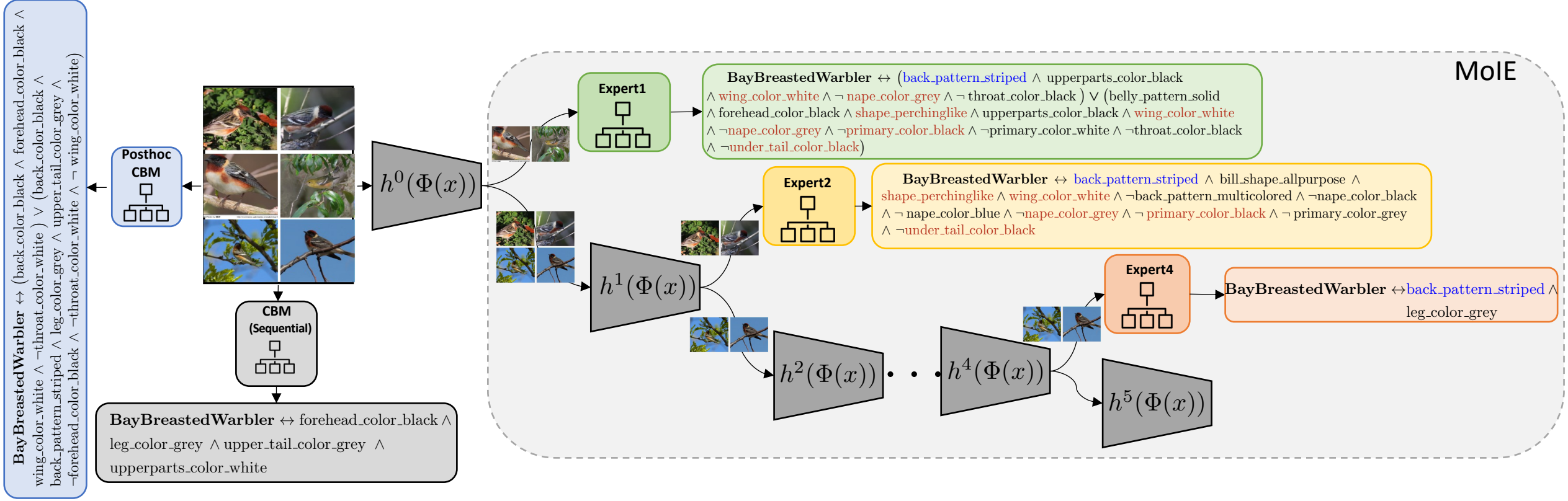


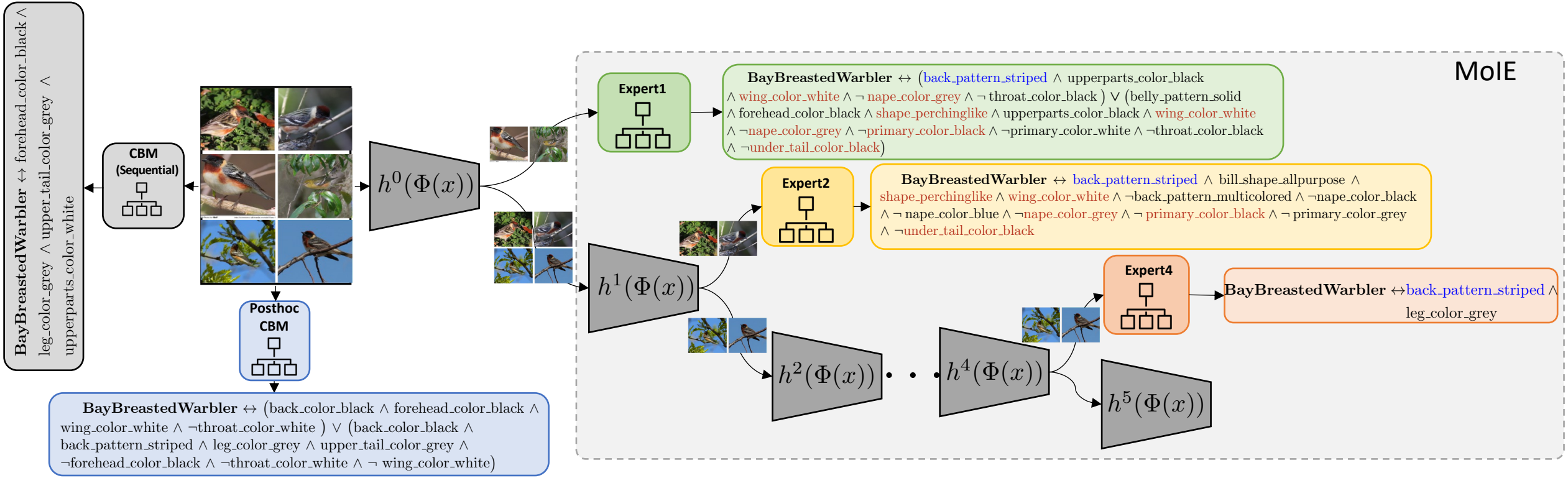
(a)

Symbol	Concept
c_1	*back_pattern_stripped*
c_2	*leg_color_grey*
c_3	*shape_perching_like*
c_4	*wing_color_white*
c_5	*nape_color_grey*
c_6	*primary_color_black*
c_7	*primary_color_grey*
c_8	*back_pattern_multicolored*
c_9	*nape_color_black*
c_{10}	*bill_shape_allpurpose*
c_{11}	*nape_color_blue*
c_{12}	*forehead_color_black*
c_{13}	*undertail_color_black*
c_{14}	*thorat_color_black*
c_{15}	*primary_color_white*
c_{16}	*belly_pattern_solid*
c_{17}	*upperparts_color_black*

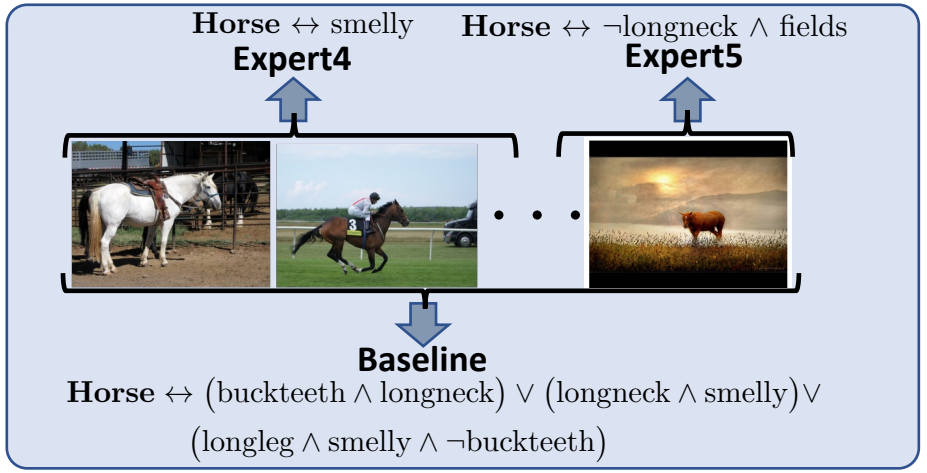
(b)

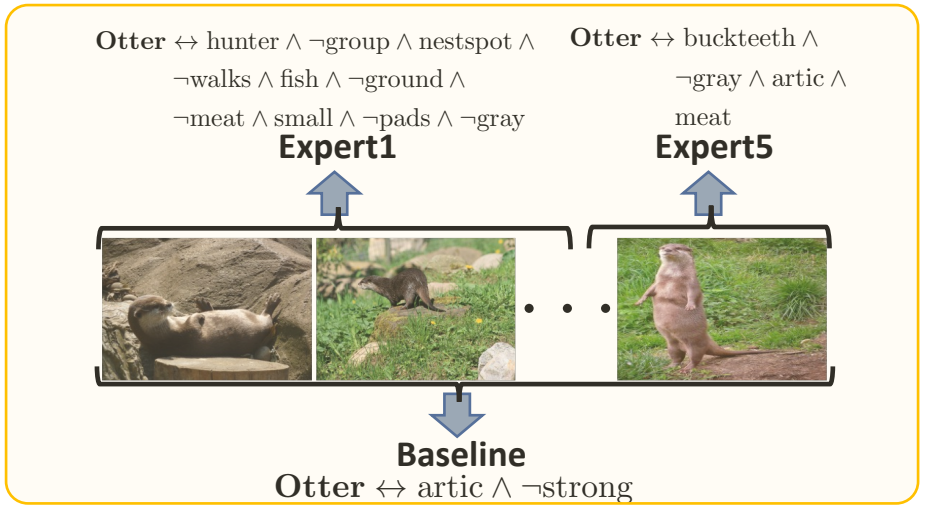

Expert1
BayBreastedWarbler ↔ (back_pattern_striped ∧ upperparts_color_black ∧ wing_color_white ∧ ¬ nape_color_grey ∧ ¬ throat_color_black) ∨ (belly_pattern_solid ∧ forehead_color_black ∧ shape_perchinglike ∧ upperparts_color_black ∧ wing_color_white ∧ ¬nape_color_grey ∧ ¬primary_color_black ∧ ¬primary_color_white ∧ ¬throat_color_black ∧ ¬under_tail_color_black)
Expert2
BayBreastedWarbler ↔ back_pattern_striped ∧ bill_shape_allpurpose ∧ shape_perchinglike ∧ wing_color_white ∧ ¬back_pattern_multicolored ∧ ¬nape_color_black ∧ ¬ nape_color_blue ∧ ¬nape_color_grey ∧ ¬ primary_color_black ∧ ¬ primary_color_grey ∧ ¬under_tail_color_black
Expert4
BayBreastedWarbler ↔ back_pattern_striped ∧ leg_color_grey
$h^0(\Phi(x))$
$h^1(\Phi(x))$
$h^2(\Phi(x))$
$h^4(\Phi(x))$
$h^5(\Phi(x))$

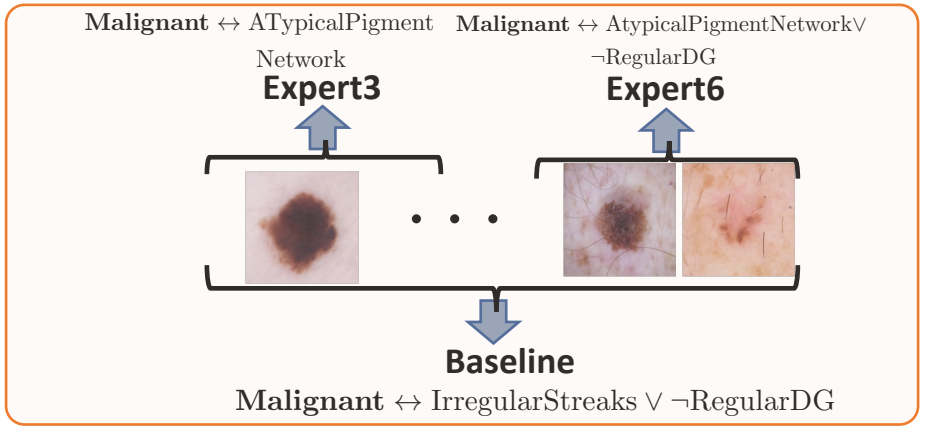

BayBreastedWarbler ↔ (back_color_black ∧ forehead_color_black ∧ wing_color_white ∧ ¬throat_color_white) ∨ (back_color_black ∧ back_pattern_striped ∧ leg_color_grey ∧ upper_tail_color_grey ∧ ¬forehead_color_black ∧ ¬throat_color_white ∧ ¬ wing_color_white)
Posthoc CBM
CBM (Sequential)
BayBreastedWarbler ↔ forehead_color_black ∧ leg_color_grey ∧ upper_tail_color_grey ∧ upperparts_color_white
$h^0(\Phi(x))$
MoIE
Expert1
BayBreastedWarbler ↔ (back_pattern_striped ∧ upperparts_color_black ∧ wing_color_white ∧ ¬ nape_color_grey ∧ ¬ throat_color_black) ∨ (belly_pattern_solid ∧ forehead_color_black ∧ shape_perchinglike ∧ upperparts_color_black ∧ wing_color_white ∧ ¬nape_color_grey ∧ ¬primary_color_black ∧ ¬primary_color_white ∧ ¬throat_color_black ∧ ¬under_tail_color_black)
Expert2
BayBreastedWarbler ↔ back_pattern_striped ∧ bill_shape_allpurpose ∧ shape_perchinglike ∧ wing_color_white ∧ ¬back_pattern_multicolored ∧ ¬nape_color_black ∧ ¬ nape_color_blue ∧ ¬nape_color_grey ∧ ¬ primary_color_black ∧ ¬ primary_color_grey ∧ ¬under_tail_color_black
$h^1(\Phi(x))$
Expert4
BayBreastedWarbler ↔back_pattern_striped ∧ leg_color_grey
$h^2(\Phi(x))$
$h^4(\Phi(x))$
$h^5(\Phi(x))$

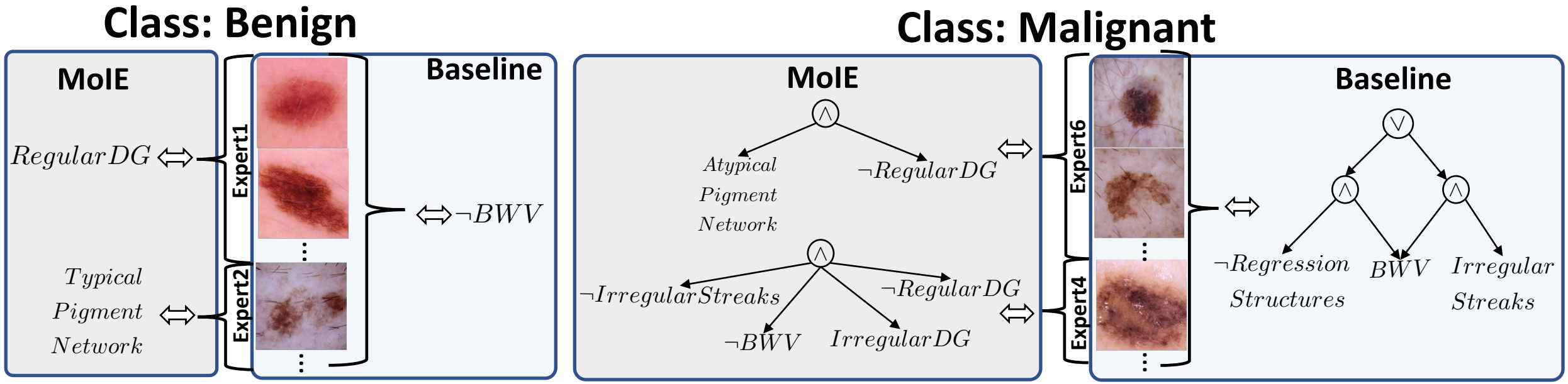


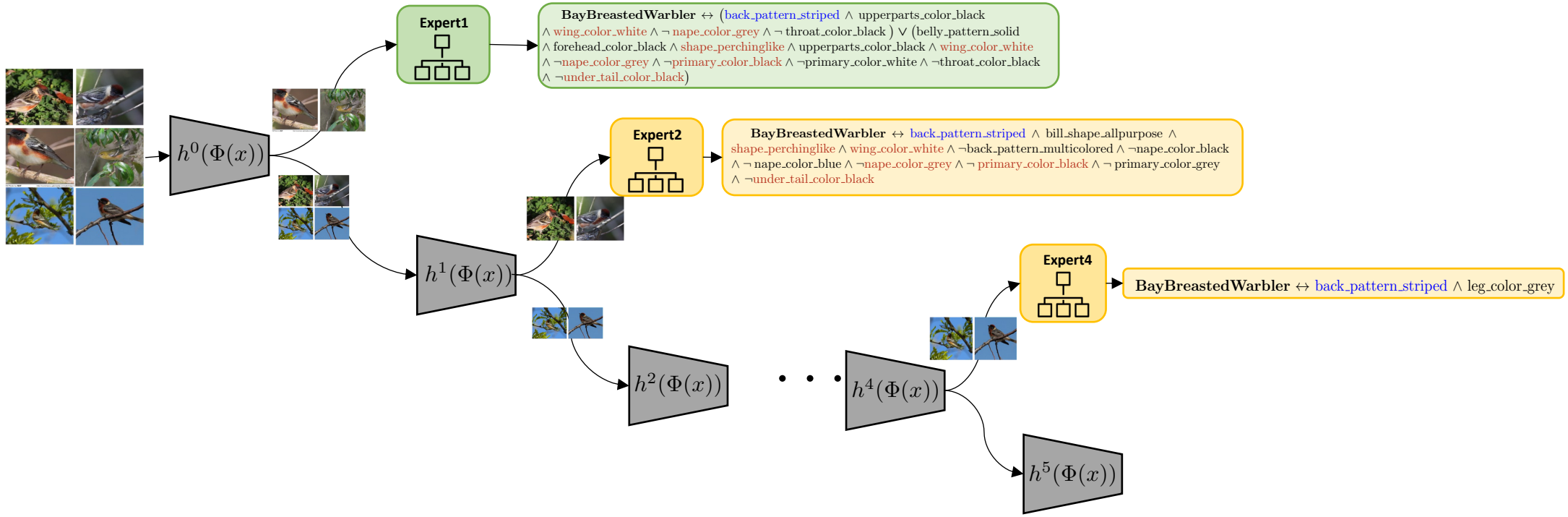
Local Logical explanations. Generic by the Baseline. Not clear which concept for which sample

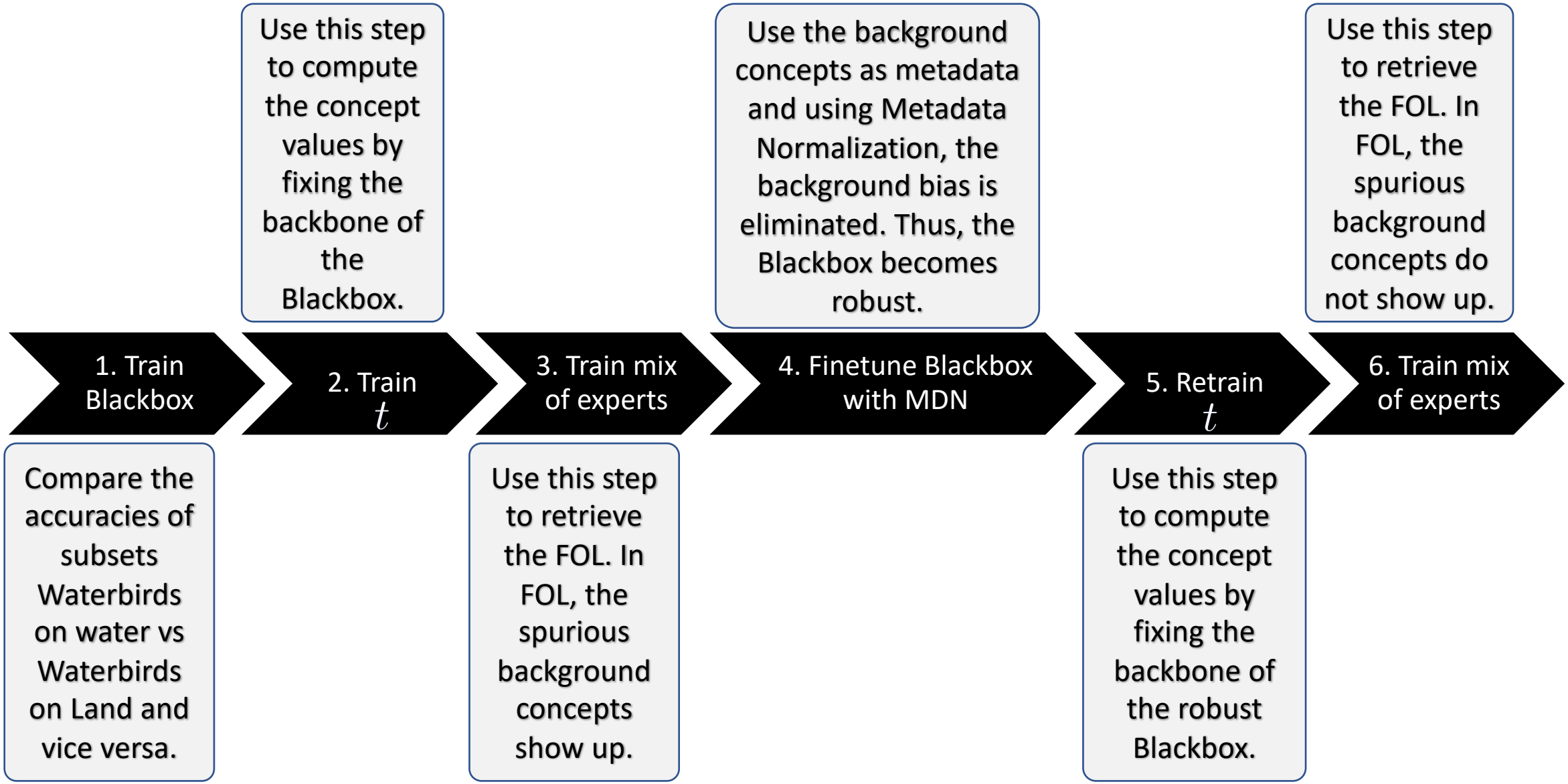

Horse ↔ smelly
Expert4
Horse ↔ ¬longneck ∧ fields
Expert5
Baseline
Horse ↔ (buckteeth ∧ longneck) ∨ (longneck ∧ smelly)∨
(longleg ∧ smelly ∧ ¬buckteeth)

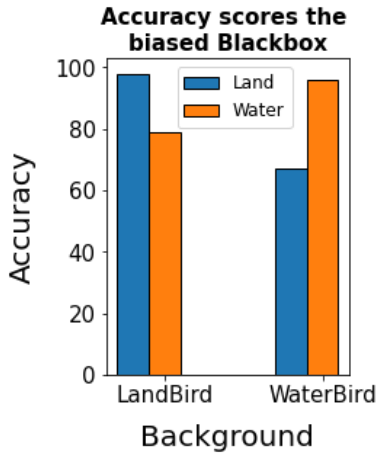

Otter ↔ hunter ∧ ¬group ∧ nestspot ∧
¬walks ∧ fish ∧ ¬ground ∧
¬meat ∧ small ∧ ¬pads ∧ ¬gray
Expert1
Otter ↔ buckteeth ∧
¬gray ∧ artic ∧
meat
Expert5
Baseline
Otter ↔ artic ∧ ¬strong


Malignant ↔ ATypicalPigment
Network
Expert3
Malignant ↔ AtypicalPigmentNetwork∨
¬RegularDG
Expert6
Baseline
Malignant ↔ IrregularStreaks ∨ ¬RegularDG

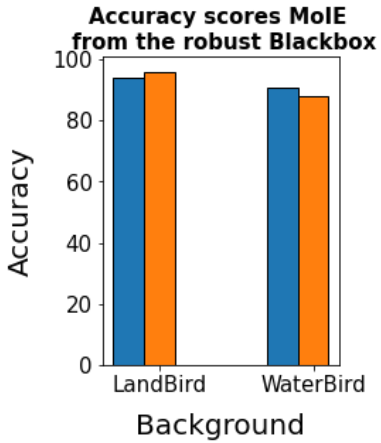

Class: Benign
MoIE
$RegularDG \Leftrightarrow$
$Typical\ Pigment\ Network \Leftrightarrow$
Expert1
Expert2
Baseline
$\Leftrightarrow \neg BWV$
(c)
Class: Malignant
MoIE
$\wedge$
$Atypical\ Pigment\ Network$
$\neg RegularDG$
$\neg IrregularStreaks$
$\neg RegularDG$
$\neg BWV$
$IrregularDG$
Expert6
Expert4
Baseline
$\vee$
$\wedge$
$\wedge$
$\neg Regression\ Structures$
BWV
$Irregular\ Streaks$
(d)


Expert1
BayBreastedWarbler ↔ (back_pattern_striped ∧ upperparts_color_black ∧ wing_color_white ∧ ¬ nape_color_grey ∧ ¬ throat_color_black) ∨ (belly_pattern_solid ∧ forehead_color_black ∧ shape_perchinglike ∧ upperparts_color_black ∧ wing_color_white ∧ ¬nape_color_grey ∧ ¬primary_color_black ∧ ¬primary_color_white ∧ ¬throat_color_black ∧ ¬under_tail_color_black)
Expert2
BayBreastedWarbler ↔ back_pattern_striped ∧ bill_shape_allpurpose ∧ shape_perchinglike ∧ wing_color_white ∧ ¬back_pattern_multicolored ∧ ¬nape_color_black ∧ ¬ nape_color_blue ∧ ¬nape_color_grey ∧ ¬ primary_color_black ∧ ¬ primary_color_grey ∧ ¬under_tail_color_black
Expert4
BayBreastedWarbler ↔ back_pattern_striped ∧ leg_color_grey
$h^0(\Phi(x))$
$h^1(\Phi(x))$
$h^2(\Phi(x))$
$h^4(\Phi(x))$
$h^5(\Phi(x))$


1. Train Blackbox
Compare the accuracies of subsets Waterbirds on water vs Waterbirds on Land and vice versa.
2. Train t
Use this step to compute the concept values by fixing the backbone of the Blackbox.
3. Train mix of experts
Use this step to retrieve the FOL. In FOL, the spurious background concepts show up.
4. Finetune Blackbox with MDN
Use the background concepts as metadata and using Metadata Normalization, the background bias is eliminated. Thus, the Blackbox becomes robust.
5. Retrain t
Use this step to compute the concept values by fixing the backbone of the robust Blackbox.
6. Train mix of experts
Use this step to retrieve the FOL. In FOL, the spurious background concepts do not show up.



(a)



(b)

Biased Blackbox

Groundtruth: WaterBird

Prediction : LandBird

Explanation : LandBird $\iff$ WingShapeRoundedwings $\wedge$ **Forest**

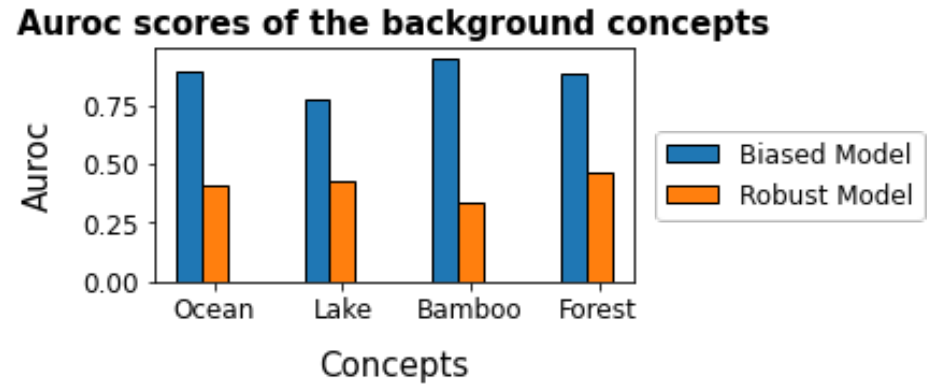
Robust Blackbox

Groundtruth: WaterBird

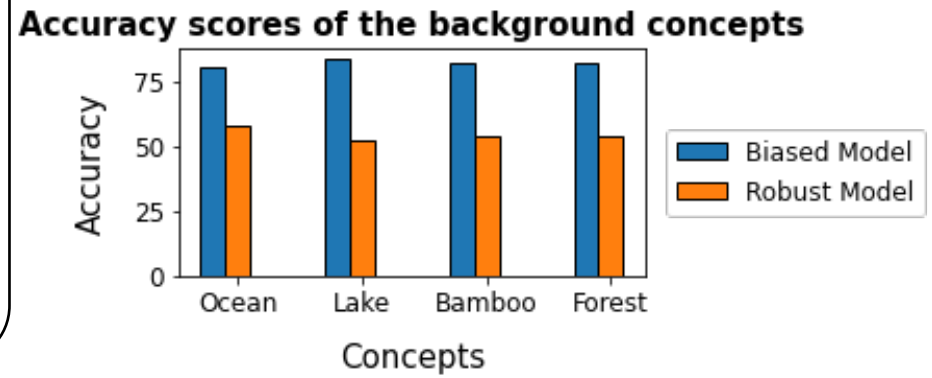
Prediction : WaterBird

Explanation : WaterBird $\iff$ BillLengthAboutTheSameAsHead$\wedge$
$\neg$BillLengthShorterThanHead $\wedge$
$\neg$SizeSmall5__9in $\wedge$ $\neg$ShapePerchingLike $\wedge$ CrownColorWhite

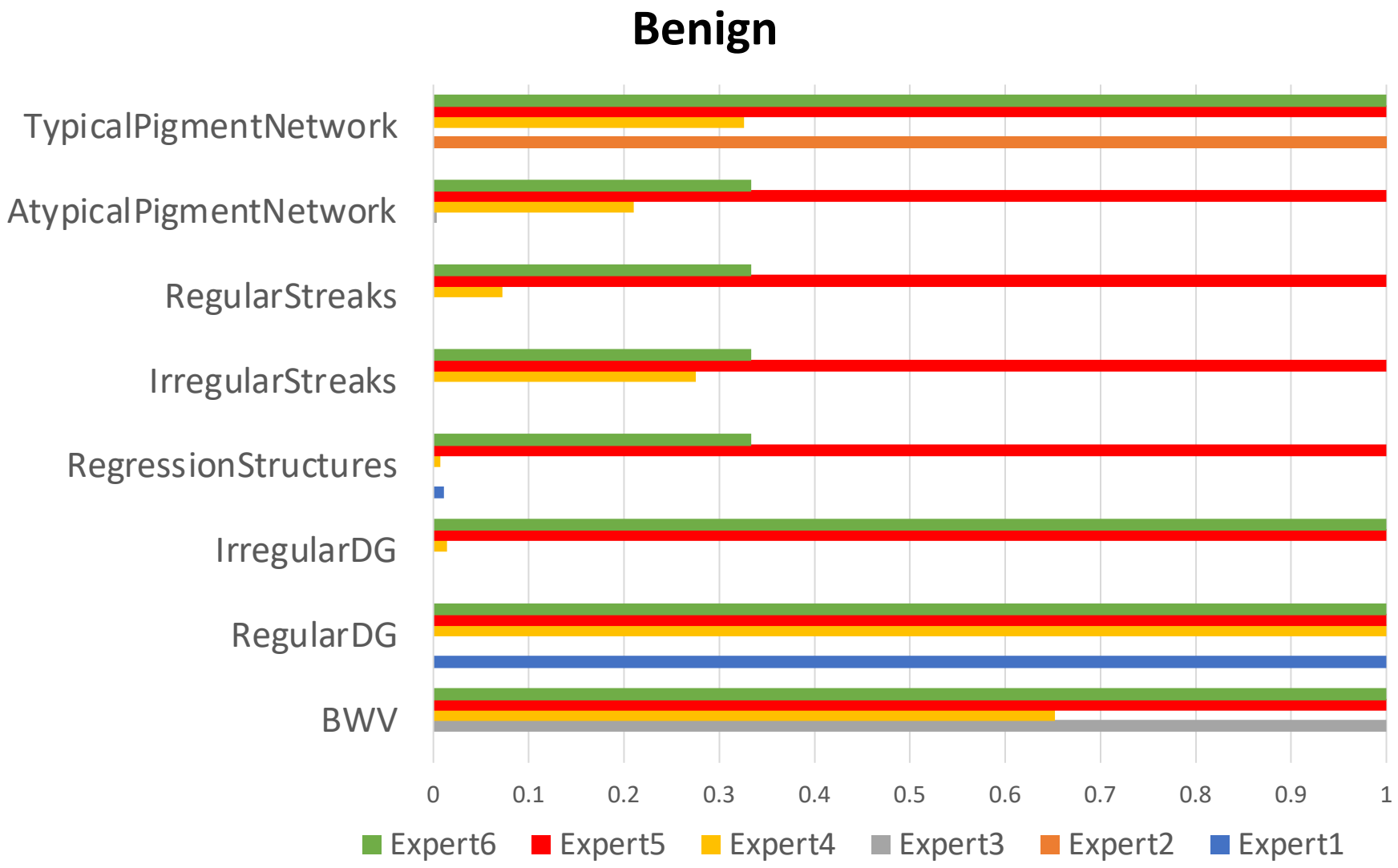
(c)

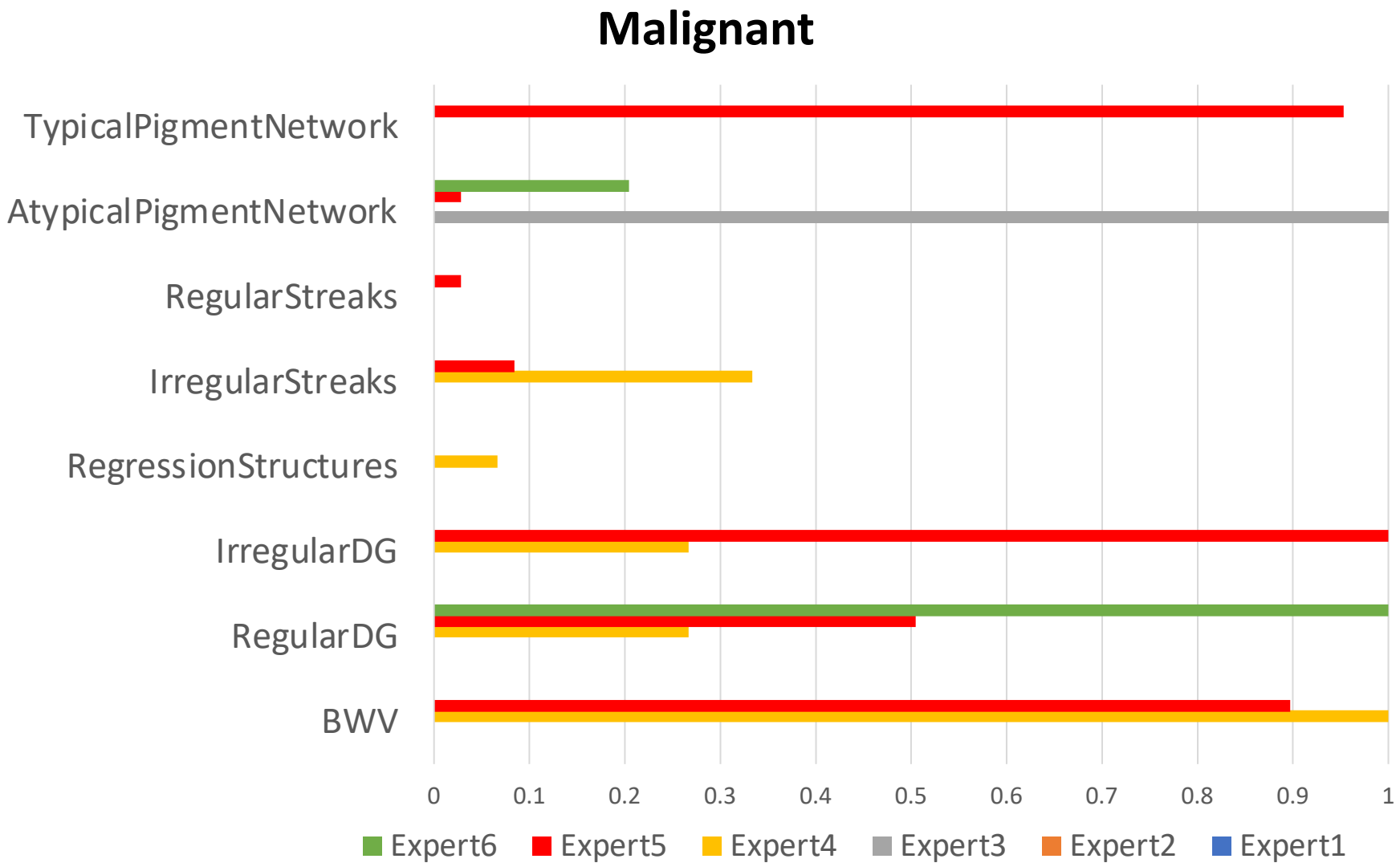


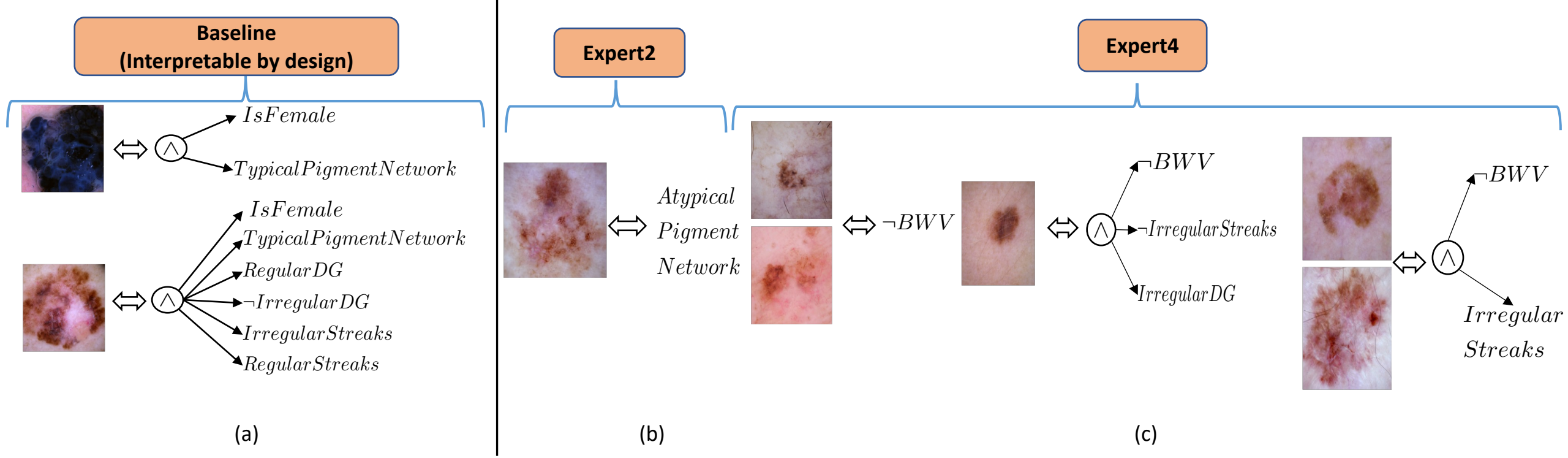
(e)

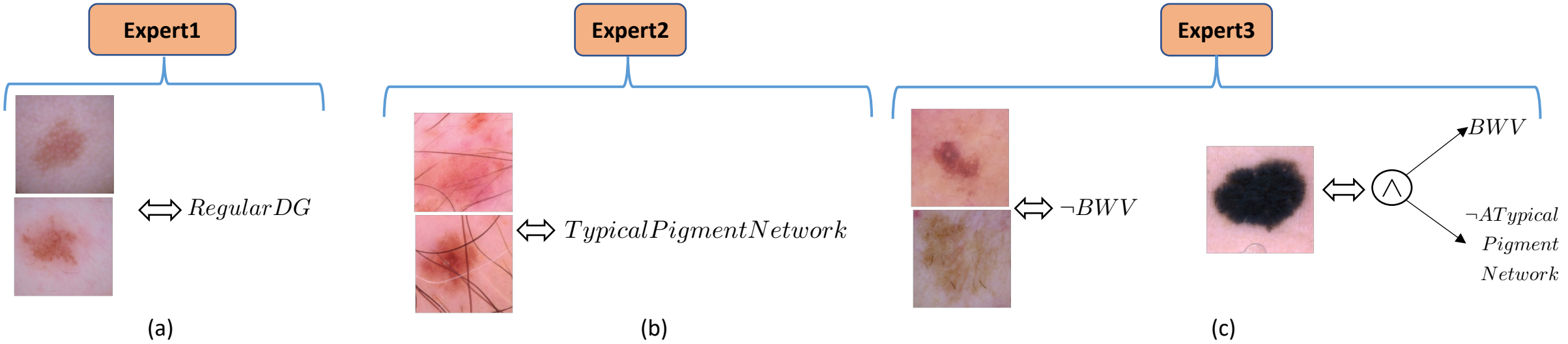


(d)

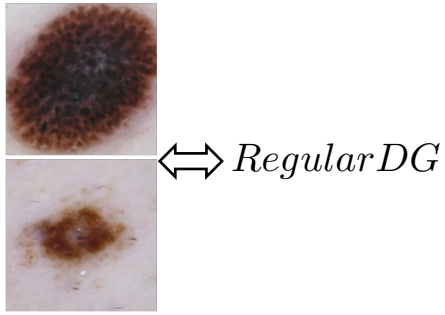

Benign
TypicalPigmentNetwork
AtypicalPigmentNetwork
RegularStreaks
IrregularStreaks
RegressionStructures
IrregularDG
RegularDG
BWV
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Expert6
Expert5
Expert4
Expert3
Expert2
Expert1


Malignant
TypicalPigmentNetwork
AtypicalPigmentNetwork
RegularStreaks
IrregularStreaks
RegressionStructures
IrregularDG
RegularDG
BWV
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Expert6
Expert5
Expert4
Expert3
Expert2
Expert1


Baseline
(Interpretable by design)
$IsFemale$
$TypicalPigmentNetwork$
$IsFemale$
$TypicalPigmentNetwork$
$RegularDG$
$\neg IrregularDG$
$IrregularStreaks$
$RegularStreaks$
(a)
Expert2
$Atypical$
$Pigment$
$Network$
Expert4
$\neg BWV$
$\neg BWV$
$\neg IrregularStreaks$
$IrregularDG$
$\neg BWV$
$Irregular$
$Streaks$
(b)
(c)

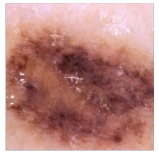

Expert1
$\Leftrightarrow RegularDG$
(a)
Expert2
$\Leftrightarrow TypicalPigmentNetwork$
(b)
Expert3
$\Leftrightarrow \neg BWV$
BWV
$\wedge$
$\neg ATypical$
$Pigment$
$Network$
(c)

Expert4

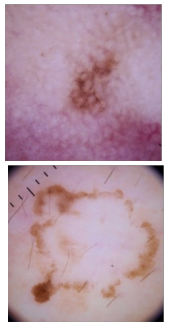


$\Leftrightarrow RegularDG$

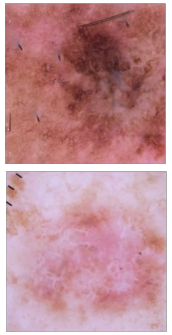
$\Leftrightarrow \wedge$ ($\neg BWV$, $\neg RegularDG$)



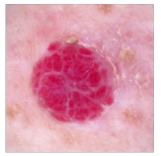
$\Leftrightarrow \wedge$ ($\neg BWV$, $\neg RegularDG$, $\neg TypicalPigmentNetwork$, $\neg IrregularStreaks$, $\neg AtypicalPigmentNetwork$, $\neg RegularStreaks$, $\neg IrregularDG$)



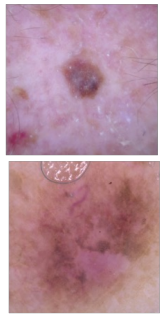
$\Leftrightarrow \wedge$ ($\neg BWV$, $\neg RegularDG$, $\neg TypicalPigmentNetwork$)



$\Leftrightarrow \wedge$ ($\neg BWV$, $\neg RegularDG$, $\neg TypicalPigmentNetwork$, $\neg IrregularStreaks$)



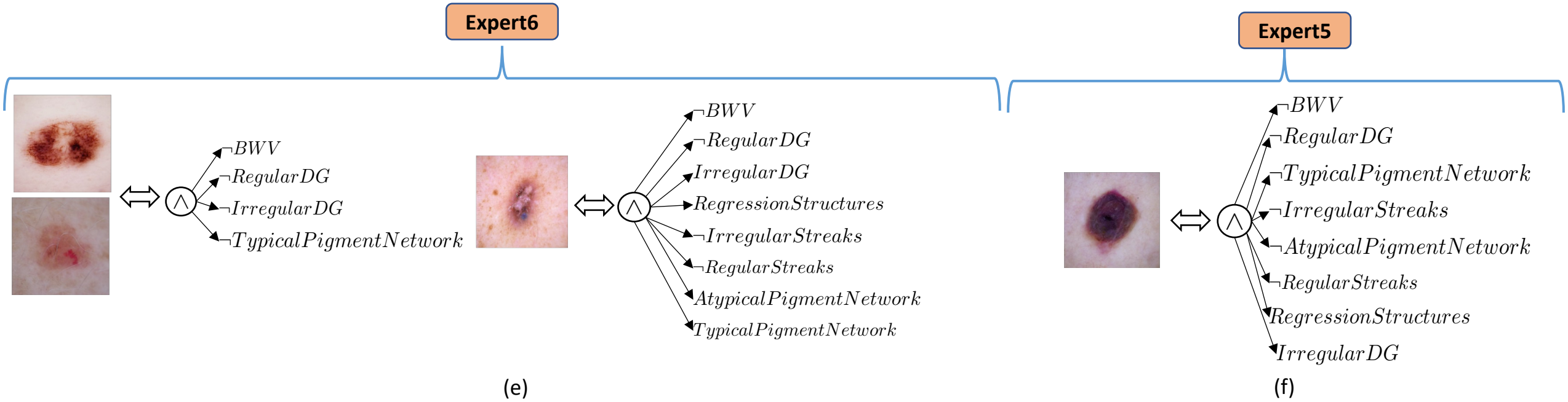
$\Leftrightarrow \wedge$ ($\neg BWV$, $\neg RegularDG$, $\neg TypicalPigmentNetwork$, $\neg IrregularStreaks$, $\neg AtypicalPigmentNetwork$)

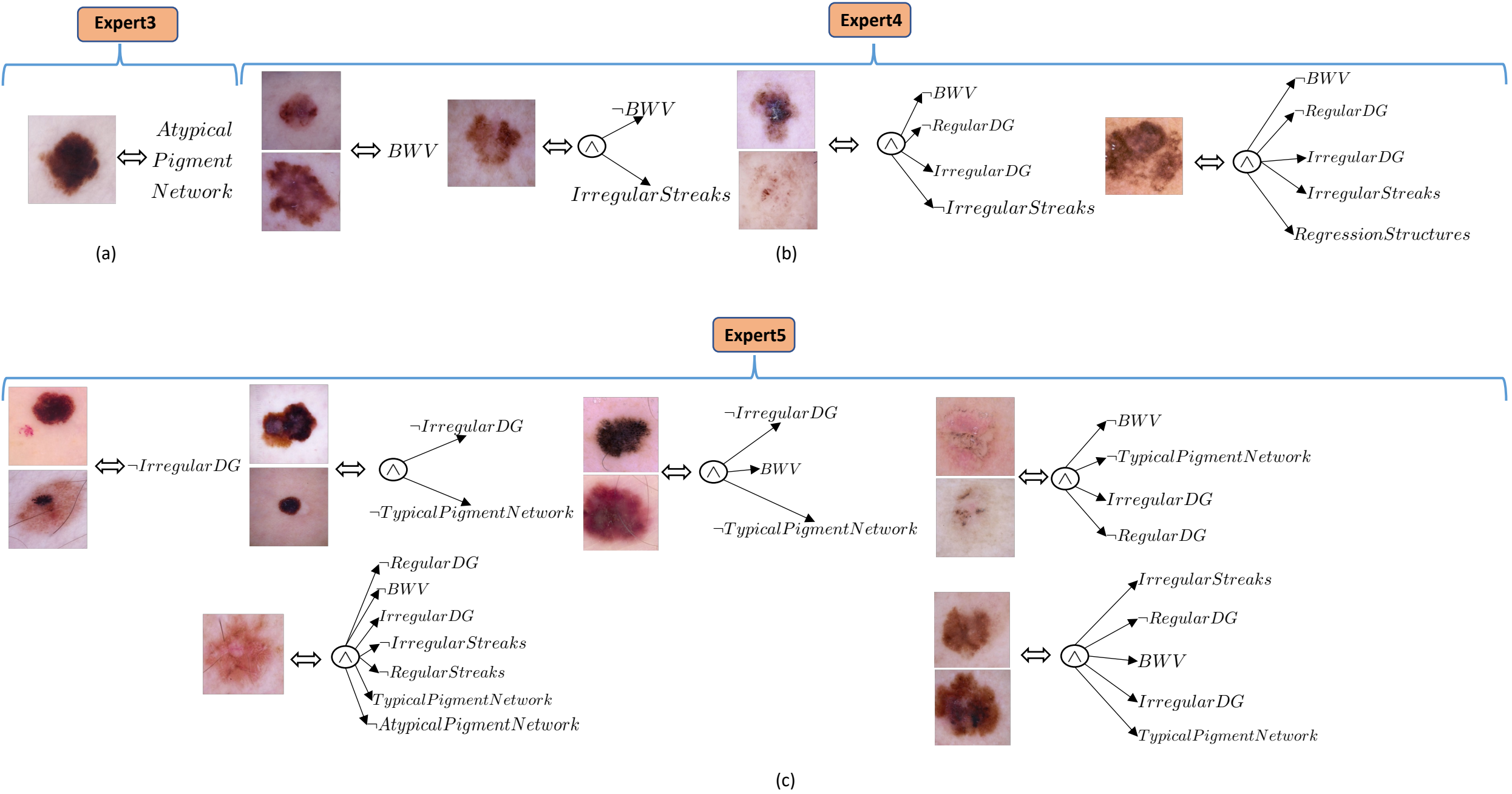


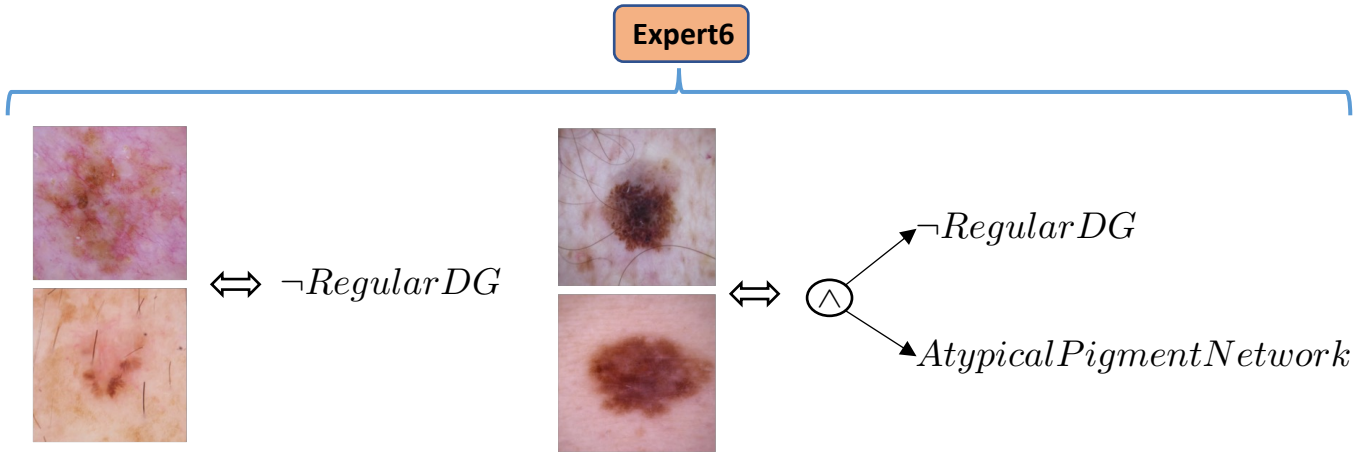
$\Leftrightarrow \wedge$ ($\neg BWV$, $\neg RegularDG$, $\neg TypicalPigmentNetwork$, $\neg IrregularStreaks$, $\neg AtypicalPigmentNetwork$, $\neg RegularStreaks$)

$\Leftrightarrow \wedge$ ($\neg BWV$, $\neg RegularDG$, $\neg TypicalPigmentNetwork$, $\neg IrregularStreaks$, $\neg AtypicalPigmentNetwork$, $\neg RegularStreaks$, $\neg RegressionStructures$, $\neg IrregularDG$)

(d)


Expert6
BWV
RegularDG
IrregularDG
TypicalPigmentNetwork
BWV
RegularDG
IrregularDG
RegressionStructures
IrregularStreaks
RegularStreaks
AtypicalPigmentNetwork
TypicalPigmentNetwork
(e)
Expert5
BWV
RegularDG
TypicalPigmentNetwork
IrregularStreaks
AtypicalPigmentNetwork
RegularStreaks
RegressionStructures
IrregularDG
(f)


Expert3
Expert4
Expert5
Atypical Pigment Network
BWV
¬BWV
IrregularStreaks
¬BWV
¬RegularDG
IrregularDG
¬IrregularStreaks
¬BWV
¬RegularDG
IrregularDG
IrregularStreaks
RegressionStructures
(a)
(b)
¬IrregularDG
¬IrregularDG
¬TypicalPigmentNetwork
¬IrregularDG
BWV
¬TypicalPigmentNetwork
¬BWV
¬TypicalPigmentNetwork
IrregularDG
¬RegularDG
¬RegularDG
BWV
IrregularDG
¬IrregularStreaks
RegularStreaks
TypicalPigmentNetwork
¬AtypicalPigmentNetwork
IrregularStreaks
¬RegularDG
BWV
IrregularDG
TypicalPigmentNetwork
(c)


Expert6
$\Leftrightarrow \neg RegularDG$
$\Leftrightarrow$
$\wedge$
$\neg RegularDG$
$AtypicalPigmentNetwork$

(d)


Expert1
Expert2
Expert3
Expert4
Expert5
Expert6
Final Residual

Painted Bunting

Expert1

$\Leftrightarrow$ $\wedge$

- $back\_pattern\_multicolored$
- leg_color_grey
- $\neg wing\_shape\_roundedwings$

Expert2

$\Leftrightarrow$ $\wedge$

- $breast\_pattern\_solid$
- leg_color_grey
- $wing\_pattern\_multicolored$
- $\neg bill\_color\_black$
- $\neg wing\_shape\_roundedwings$

Baltimore Oriole

Expert1

$\Leftrightarrow$ $\wedge$

- $\neg wing\_shape\_roundedwings$
- $back\_pattern\_multicolored$
- $bill\_color\_grey$
- $head\_pattern\_plain$

Expert3

$\Leftrightarrow$ $\vee$

- $\wedge$
  - $wing\_color\_white$
  - $bill\_color\_grey$
  - $head\_pattern\_plain$
  - $back\_pattern\_multicolored$
- $\wedge$
  - $back\_pattern\_multicolored$
  - $\neg breast\_pattern\_solid$
  - $\neg bill\_color\_grey$
  - $\neg wing\_color\_white$

American Goldfinch

Expert1

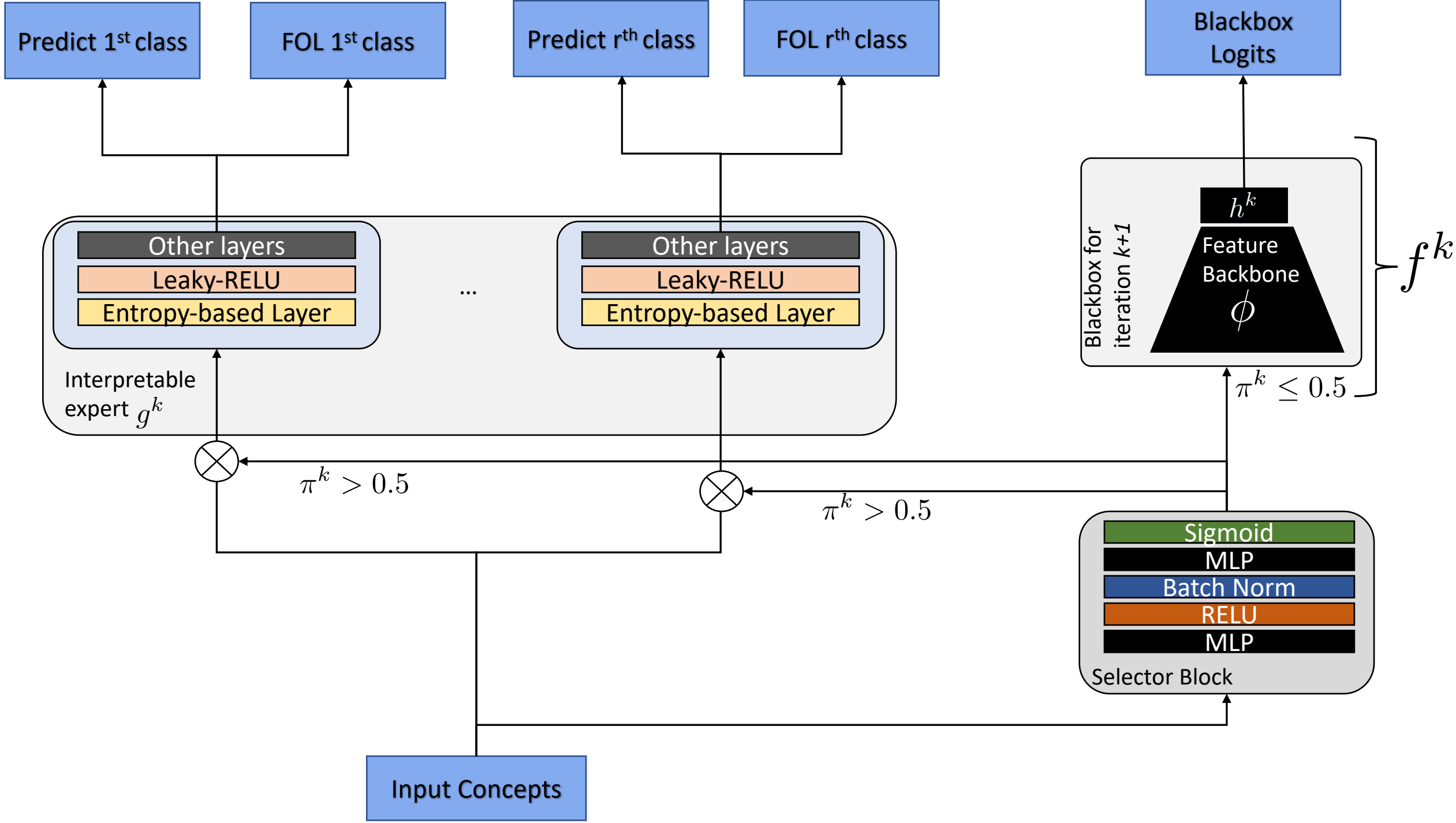
$\Leftrightarrow$ $\wedge$

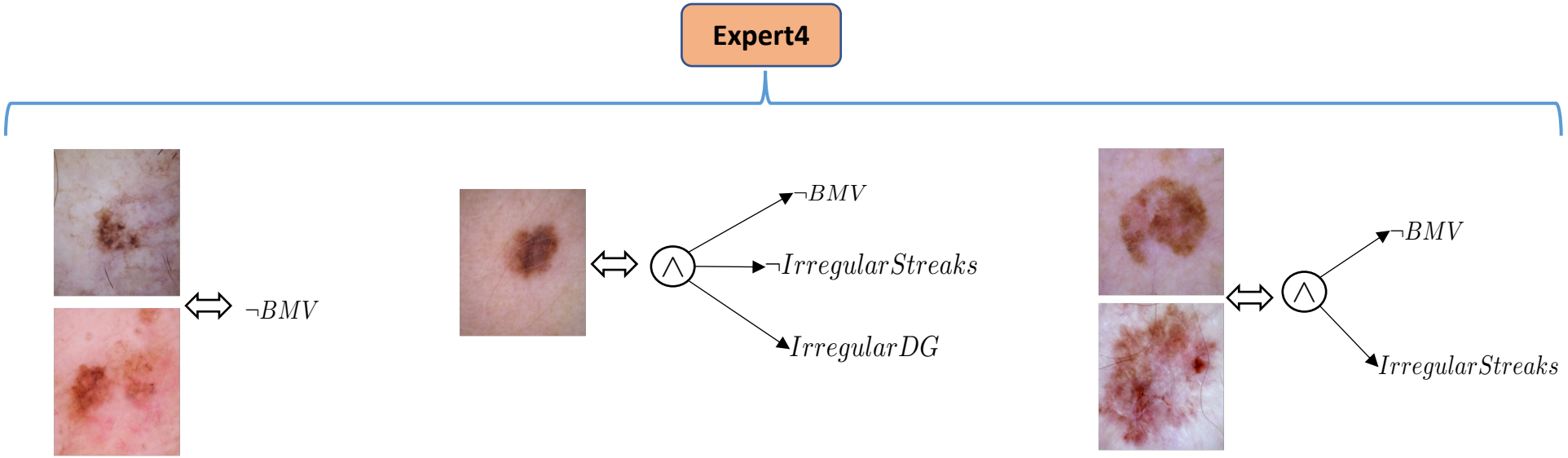
- $under\_tail\_color\_white$
- $bill\_shape\_cone$
- $upperparts\_color\_yellow$
- $upper\_tail\_color\_white$
- $back\_pattern\_solid$
- $\neg bill\_color\_black$
- $\neg forehead\_color\_yellow$
- $\neg leg\_color\_grey$
- $\neg size\_medium\_9\_16\_in$

Expert2

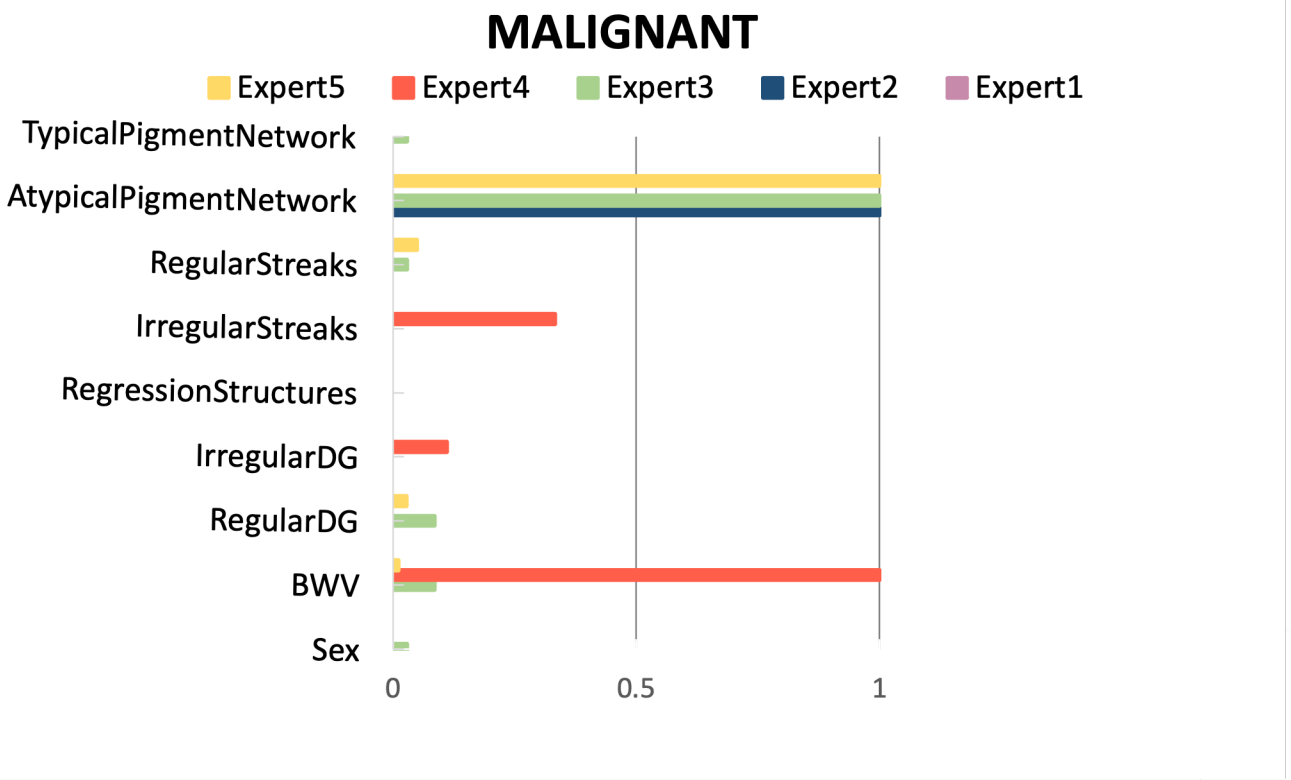
$\Leftrightarrow$ $\wedge$

- $belly\_color\_yellow$
- $upper\_tail\_color\_white$
- $wing\_pattern\_multicolored$
- $upperparts\_color\_yellow$
- $crown\_color\_yellow$
- $bill\_shape\_cone$

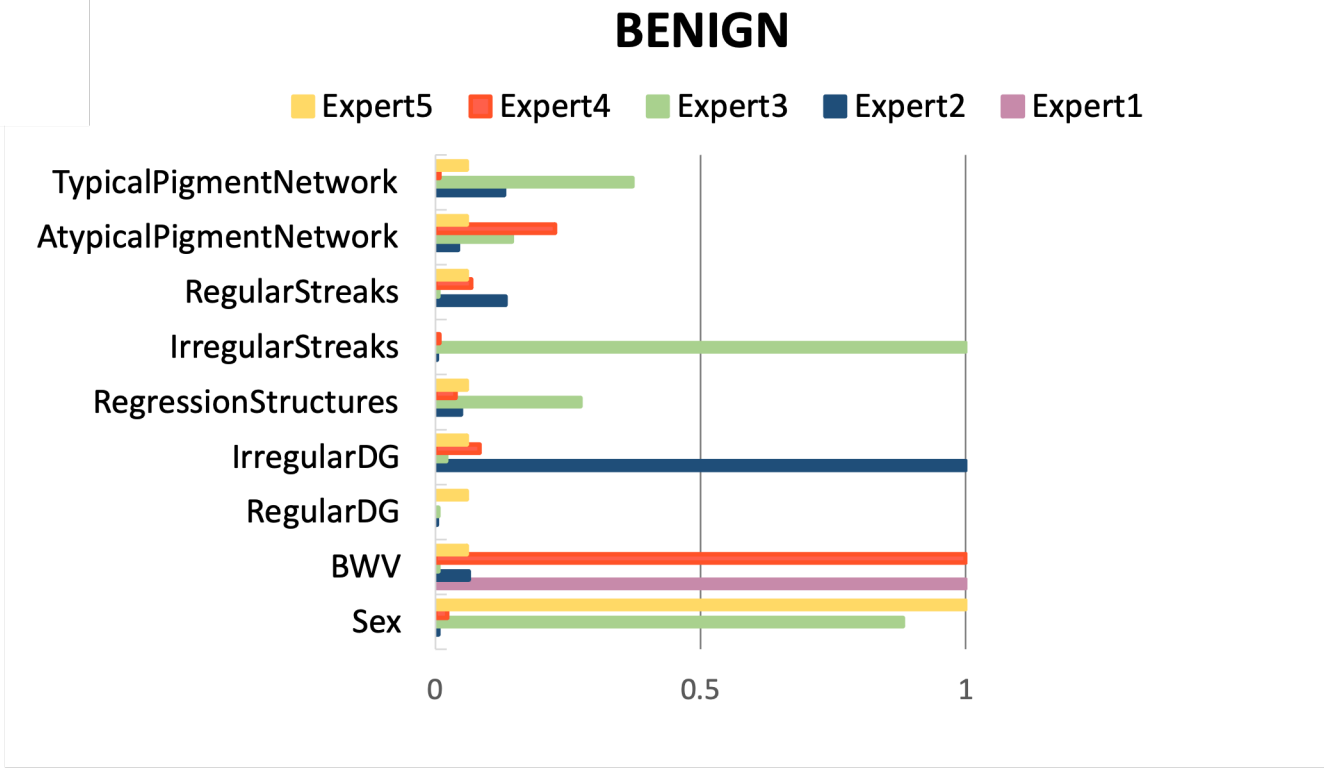

Predict 1st class
FOL 1st class
Predict rth class
FOL rth class
Blackbox Logits
Other layers
Leaky-RELU
Entropy-based Layer
...
Other layers
Leaky-RELU
Entropy-based Layer
Interpretable expert g^k
Blackbox for iteration k+1
h^k
Feature Backbone
ϕ
f^k
$\pi^k \leq 0.5$
$\pi^k > 0.5$
$\pi^k > 0.5$
Sigmoid
MLP
Batch Norm
RELU
MLP
Selector Block
Input Concepts

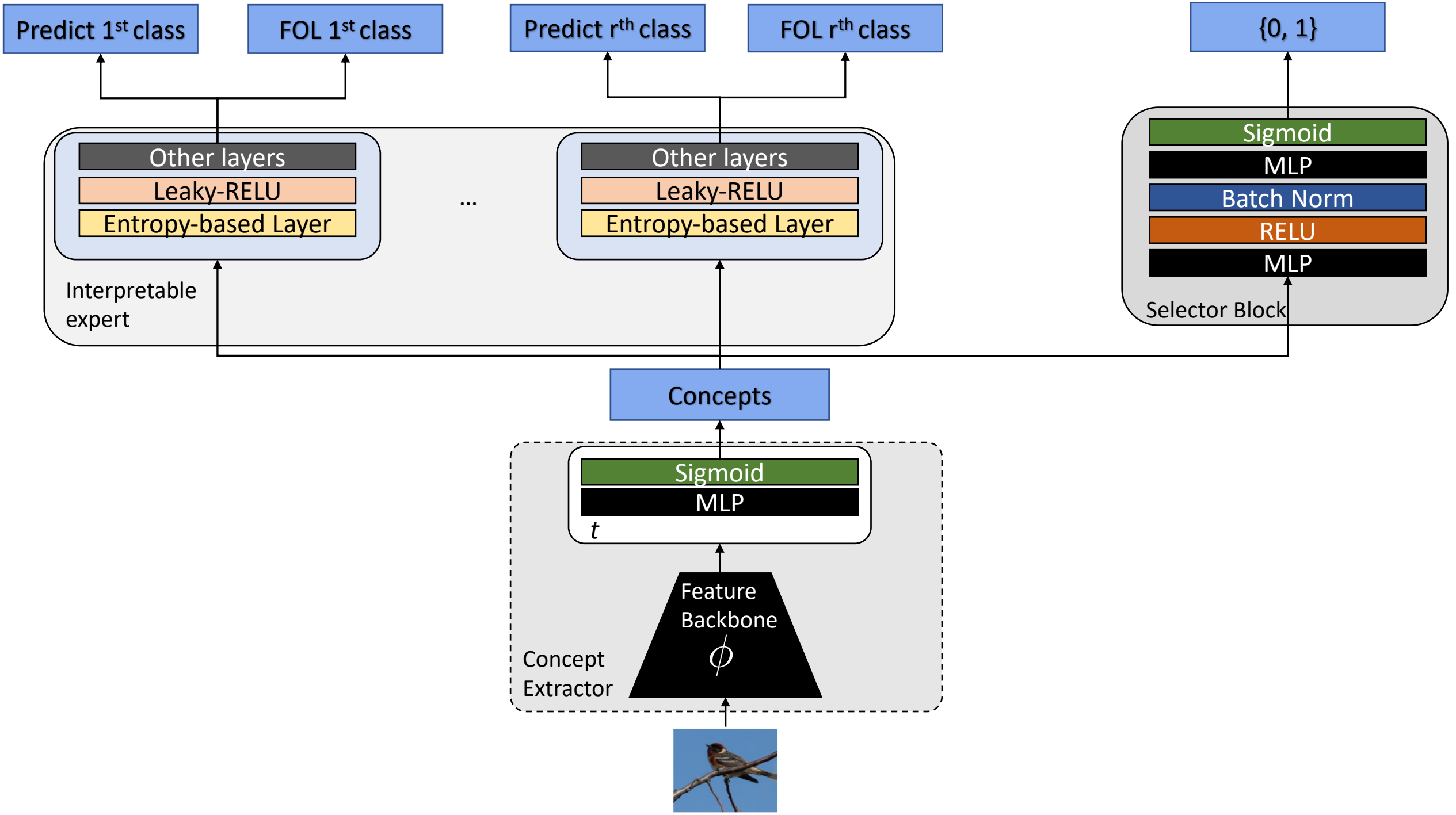

Expert4
$\neg BMV$
$\neg BMV$
$\neg IrregularStreaks$
$IrregularDG$
$\neg BMV$
$IrregularStreaks$

(c)

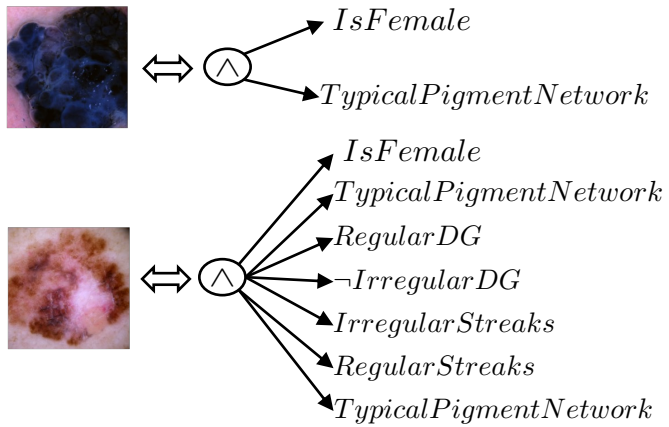

MALIGNANT
Expert5
Expert4
Expert3
Expert2
Expert1
TypicalPigmentNetwork
AtypicalPigmentNetwork
RegularStreaks
IrregularStreaks
RegressionStructures
IrregularDG
RegularDG
BWV
Sex
0
0.5
1

(a)


BENIGN
Expert5
Expert4
Expert3
Expert2
Expert1
TypicalPigmentNetwork
AtypicalPigmentNetwork
RegularStreaks
IrregularStreaks
RegressionStructures
IrregularDG
RegularDG
BWV
Sex
0
0.5
1

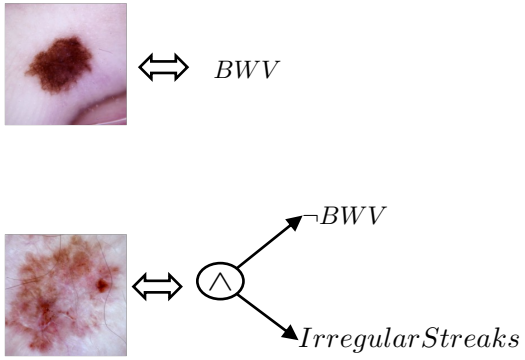

Predict 1st class
FOL 1st class
Predict rth class
FOL rth class
{0, 1}
Other layers
Leaky-RELU
Entropy-based Layer
...
Other layers
Leaky-RELU
Entropy-based Layer
Interpretable expert
Sigmoid
MLP
Batch Norm
RELU
MLP
Selector Block
Concepts
Sigmoid
MLP
t
Feature Backbone
ϕ
Concept Extractor

Interpretable by design



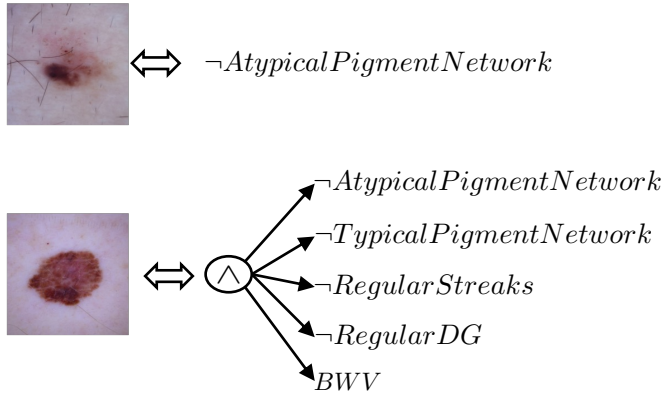
(a)

Expert4

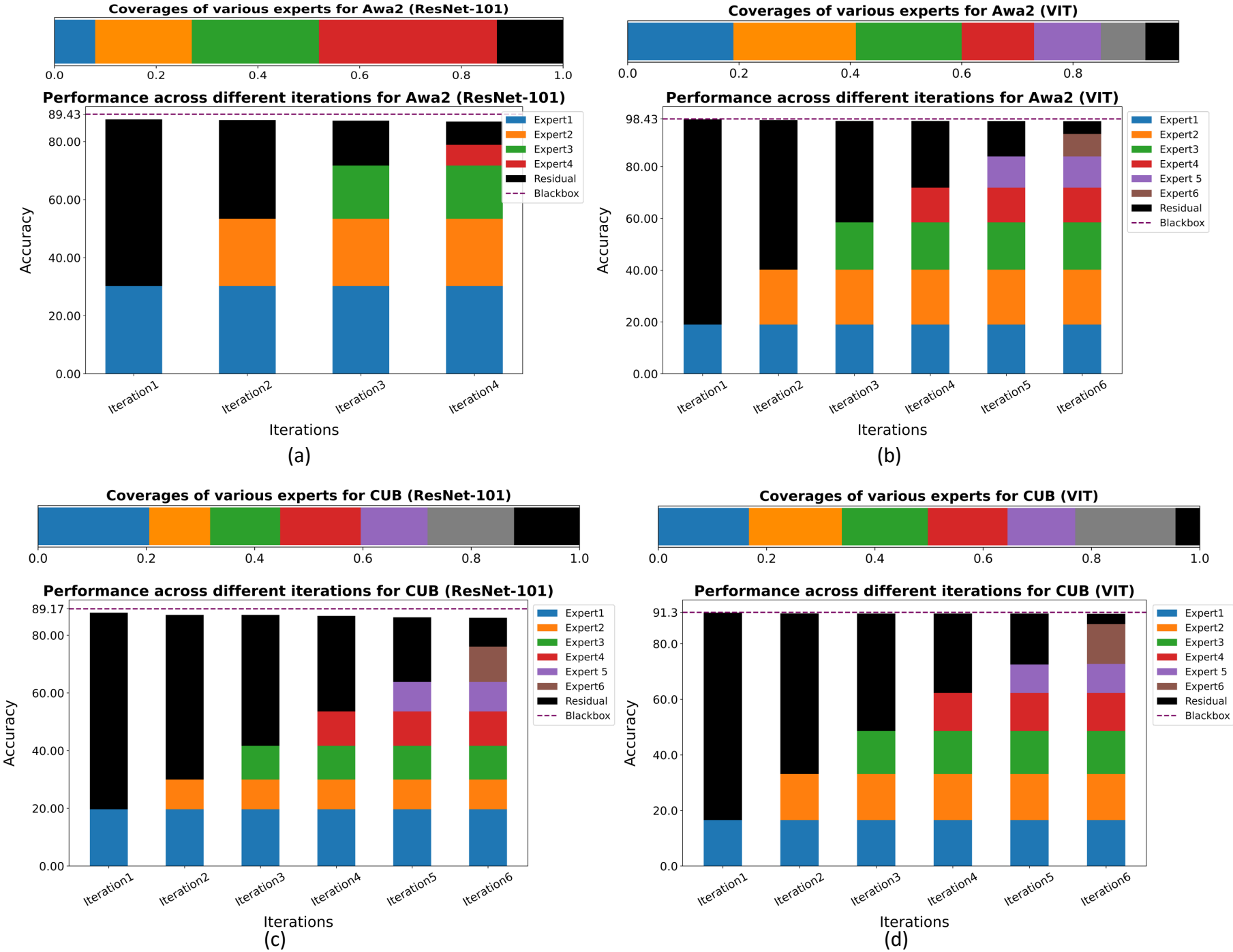


(b)

Expert5



(c)


Coverages of various experts for Awa2 (ResNet-101)
0.0
0.2
0.4
0.6
0.8
1.0
Performance across different iterations for Awa2 (ResNet-101)
89.43
80.00
60.00
40.00
20.00
0.00
Accuracy
Expert1
Expert2
Expert3
Expert4
Residual
Blackbox
Iteration1
Iteration2
Iteration3
Iteration4
Iterations
(a)
Coverages of various experts for Awa2 (VIT)
0.0
0.2
0.4
0.6
0.8
Performance across different iterations for Awa2 (VIT)
98.43
80.00
60.00
40.00
20.00
0.00
Accuracy
Expert1
Expert2
Expert3
Expert4
Expert 5
Expert6
Residual
Blackbox
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations
(b)
Coverages of various experts for CUB (ResNet-101)
0.0
0.2
0.4
0.6
0.8
1.0
Performance across different iterations for CUB (ResNet-101)
89.17
80.00
60.00
40.00
20.00
0.00
Accuracy
Expert1
Expert2
Expert3
Expert4
Expert 5
Expert6
Residual
Blackbox
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations
(c)
Coverages of various experts for CUB (VIT)
0.0
0.2
0.4
0.6
0.8
1.0
Performance across different iterations for CUB (VIT)
91.3
80.0
60.0
40.0
20.0
0.0
Accuracy
Expert1
Expert2
Expert3
Expert4
Expert 5
Expert6
Residual
Blackbox
Iteration1
Iteration2
Iteration3
Iteration4
Iteration5
Iteration6
Iterations
(d)

Seal

Expert1

$\Leftrightarrow$ $\wedge$ $\neg lean$ $\neg strainteeth$ $small$

Expert2

$\Leftrightarrow$ $\wedge$ $\neg lean$ $\neg chewteeth$ $spots$ $small$

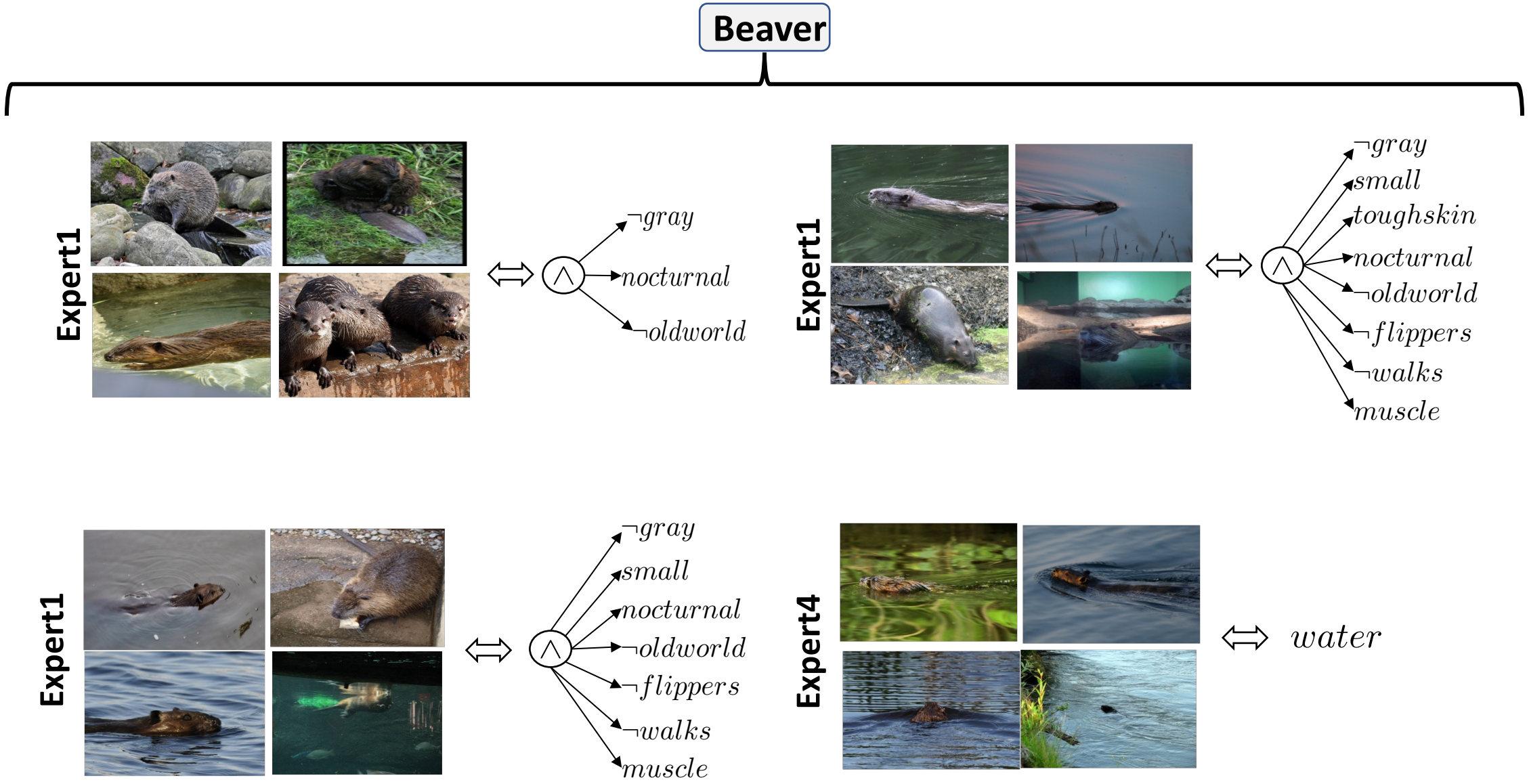
Horse

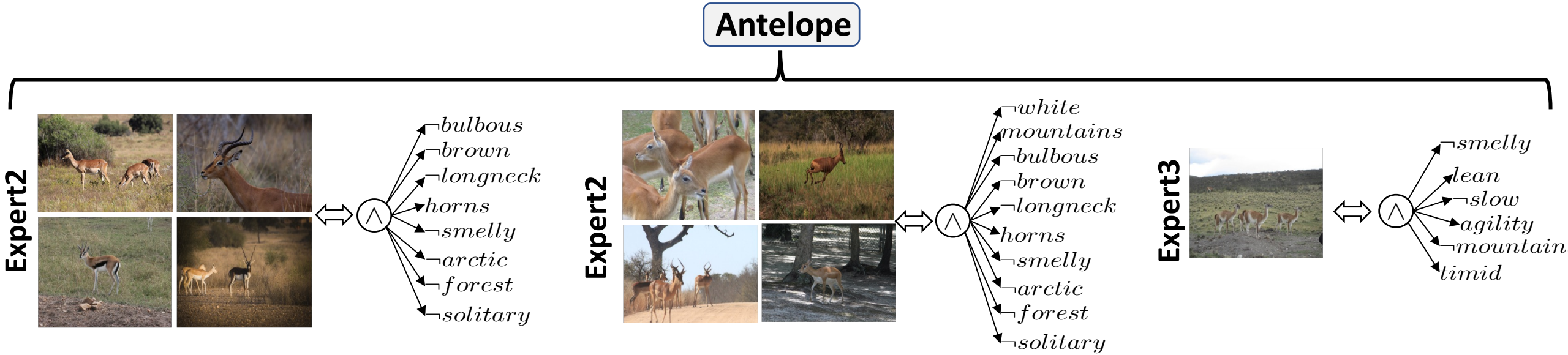
Expert4

$\Leftrightarrow$ $smelly$

Expert5

$\Leftrightarrow$ $\wedge$ $\neg longneck$ $fields$


Beaver
Expert1
¬gray
nocturnal
¬oldworld
Expert1
¬gray
small
toughskin
nocturnal
¬oldworld
¬flippers
¬walks
muscle
Expert1
¬gray
small
nocturnal
¬oldworld
¬flippers
¬walks
muscle
Expert4
water


Antelope
Expert2
¬bulbous
¬brown
¬longneck
¬horns
¬smelly
¬arctic
¬forest
¬solitary
Expert2
¬white
¬mountains
¬bulbous
¬brown
¬longneck
¬horns
¬smelly
¬arctic
¬forest
¬solitary
Expert3
¬smelly
¬lean
¬slow
¬agility
¬mountain
¬timid